

Climate of distrust

Six months into President George W. Bush's second term of office, partisan politics continues to widen the gulf between researchers and the administration.

The story has become so familiar that new twists in the plot cease to outrage. Time after time, in agency after agency, political factors have prevented US science from serving its time-honoured role in informing government decisions.

In one of the latest examples, Congressman Joe Barton (Republican, Texas) has asked three climate researchers, along with the heads of the National Science Foundation and the Intergovernmental Panel on Climate Change (IPCC), for background information on certain palaeoclimate research. On the surface, such a request seems natural. The congressman heads the House of Representatives' Committee on Energy and Commerce, which deals with environmental matters, among other topics.

But dig a little deeper and you find Joe Barton is not known as a friend of the environment. His home town's newspaper, *The Dallas Morning News*, nicknamed him 'Smokey Joe' for his efforts to exempt cement plants in his district from stricter anti-smog rules. This spring he helped the House pass an energy bill that contains no measures at all to limit greenhouse-gas emissions.

Flaws and errors

As a congressman, Barton is well within his rights to request information from researchers funded by US taxpayers. His letters to the climate scientists are not a subpoena and do not legally require a response. But their tone leaves no doubt as to his agenda.

"Questions have been raised, according to a February 14, 2005 article in *The Wall Street Journal*," he writes, "about the significance of methodological flaws and data errors in your studies of the historical record of temperatures and climate change... We open this review because this dispute surrounding your studies bears directly on important questions about the federally funded work upon which climate studies rely and the quality and transparency of analyses used to support the IPCC assessment process."

The letters request information on the scientists' professional background, financial backing, accessibility of their data and computer codes, relationship to the IPCC, and more. The requests have their genesis in a 1998 *Nature* article (M. E. Mann *et al.* *Nature* 392, 779–787; 1998), which showed that temperatures in the Northern Hemisphere rose sharply in the twentieth century, in a graphical upswing dubbed the 'hockey stick'. Michael Mann of the University of Virginia, Raymond Bradley of the University of Massachusetts, and Malcolm Hughes of the University of Arizona have spent the past few years responding to queries about this work from Canadian researchers Steve McIntyre, who worked in the mining industry, and Ross McKittrick, an economist at the University of Guelph, Ontario.

Subsequent studies have supported the observation that unprecedented warming occurred in the 20th century, while multiple lines of evidence support the notion that anthropogenic influences are contributing to it. Science is, by its very nature, a process open to the

questioning and overthrowing of currently accepted ideas, and the detail of Mann and colleagues' work has itself been debated within the climate community. Barton, in contrast, has chosen to cherry-pick selected information on the hockey-stick work, using an article from *The Wall Street Journal* as his scientific guide.

Even President George Bush, in widely reported comments last weekend, now accepts that humans are contributing to climate change. But by requesting information on research that does not fit his world view, Barton seems determined to use his political influence to put pressure on the scientific process.

Questions remain over how this might affect scientists' contributions to the next IPCC report, due out in 2007 but being written now. Climate researchers will recall the case of Benjamin Santer, an IPCC contributor who took a beating from climate sceptics over allegedly altering the 1995 IPCC report to play up the anthropogenic influence on climate. Some wonder whether Mann, whose work played a prominent role in the 2001 report, may be facing pressure for similar reasons.

Rajendra Pachauri, chairman of the IPCC, may be correct that, overall, Barton's letters are not a threat to the scientific community or the integrity of the IPCC review process (see page 7), but there is no room for complacency. After all, other areas of US science have not been so lucky. Federal reports on climate research have been altered, in part by former oil and gas lobbyists, to play down human effects on climate. Officials at the Environmental Protection Agency have been busy removing key portions from reports or altering conclusions on the orders of political appointees. And a recent survey of National Oceanic and Atmospheric Administration scientists revealed that nearly a third felt they could not do their jobs properly in the face of interference from non-scientist administrators.

Roughed up

Although Jack Marburger, President Bush's science adviser and director of the White House Office of Science and Technology Policy, continues to defend the administration's record, some US congressmen have taken up the banner of science. Representative Henry Waxman (Democrat, California) has asked Barton to withdraw the letters and instead hold a hearing on climate change, perhaps inviting the letters' recipients as witnesses.

The politicization of science in the United States has deepened since last November's election. But the US research community cannot simply wait for things to improve in three and a half years' time. They must speak out on each contentious issue and ensure that the genuine voices of science are heard.

"Congressman Barton has chosen to cherry-pick selected information, using an article from *The Wall Street Journal* as his guide."



Rules of engagement

Biologists may soon have little option but to sign up to codes of conduct.

Why don't people resent driving licences? After all, most drivers don't intend to go out and kill people and they don't like their freedom to roam restricted. But road vehicles can be turned into a means of destruction, whether through carelessness, mischief or malice. People accept, therefore, a licence that reinforces rules designed to help keep people safe.

Substitute 'organisms' for 'road vehicles' and 'researchers' for 'drivers' and you have a case for the licensing of biologists. And however outrageous *Nature* readers may consider it, politicians and policy-makers are taking codes of conduct and licensing in research seriously. This much was clear from last week's inaugural meeting of the US National Science Advisory Board for Biosecurity (NSABB). Just a US problem? Applicable only to defence labs? Think again. Any biologists whose work could be misused (and they are many) face growing pressure to reassure their fellow citizens. This would also pre-empt excessive regulation by a willingness to declare explicitly just how responsible they intend to be.

Most researchers would wonder what planet such proposals come from. Speaking at the NSABB meeting, a social scientist, Brian Rappert of Exeter University, UK, described discussions with 600 biologists in the United Kingdom, most of whom were blissfully unaware of issues of 'dual use' that learned societies, biological weapons convention negotiators and others have been fretting about for years. Speakers from the intelligence services and weapons inspectorates emphasized, however, the active interest in the harmful uses of biology they have encountered among a malevolent few.

The approach favoured at the NSABB's preliminary discussions

was to all but accept that genes cannot be kept in bottles, and to focus instead on developing a 'culture of responsibility' in research institutions, journals and, especially, among individuals (see *Nature* 435, 860; 2005). Engineers and medics are imbued with a culture that, if ignored, can lead to professional prohibition. For most basic researchers, codes of behaviour, although taken for granted where safety and human or animal experimentation are concerned, are alien when it comes to broader professional practice.

Consider the phrase 'do no harm'. Deceptively simple, a trite piece of motherhood and apple pie, and yet, as one medical researcher at the meeting said, this fundamental principle had provided him with significant help when faced with some critical professional decisions.

As Rappert highlighted, many codes for biologists have been drafted, ranging from such statements of aspiration to enforceable codes of practice (see www.projects.ex.ac.uk/codesofconduct).

If they are to carry any weight, they will need to be backed up by certification, and research institutions will need to extend their often flimsy means of assuring compliance. Less demandingly, good behaviour can be encouraged by education, as illustrated at the meeting by Nobel laureate Phillip Sharp with examples of courses on science ethics and best practice at the Massachusetts Institute of Technology.

Codes of practice have so far attracted little attention in the biology community. But in a world threatened by terrorism, governments are taking more interest in such codes, and scientists would do well to engage in a constructive discussion about what role they might play. ■

"If codes are to carry any weight they will need to be backed by certification, and institutions will have to extend their means of assuring compliance."

Playing the name game

Stem-cell biologists should not try to change the definition of the word 'embryo'.

Last month's meeting of the International Society for Stem Cell Research in San Francisco witnessed a bizarre semantic debate. Delegates discussed a proposal to refrain from using the term 'embryo' when referring to the blastocysts from which human embryonic stem cells are harvested. The scientists involved reject the accusation that they are creating and destroying human lives, and fear that the word 'embryo' is a lightning rod that attracts negative scrutiny.

It is true that embryo is an emotive term, but there is little scientific justification for redefining it. Whether taken from a fertility clinic or made through cloning, a blastocyst embryo has the potential to become a fully functional organism. And appearing to deny that fact will not fool die-hard opponents of this research. If anything, it will simply open up scientists to the accusation that they are trying to distance themselves from difficult moral

issues by changing the terms of the debate.

At the equivalent meeting last year, the society decided to formally adopt the term 'somatic cell nuclear transfer' to describe the procedure in which an adult cell nucleus is transplanted into an egg to produce embryonic stem cells. This procedure had been called 'therapeutic cloning' to distinguish it from 'reproductive cloning', which would use the same technique in an attempt to make a baby.

But the work is far from yielding any therapies, and scientists realized that the word 'cloning' was generating public concern. So they decided to adopt a more technical term less likely to stir up strong emotions. At least that re-branding had the positive effect of toning down the hype surrounding therapeutic cloning.

The name change debated at last month's meeting would be a step too far, however. In the future, researchers may isolate pluripotent stem cells from biological entities that do not have the same developmental potential as embryos. This may justify the creation of a new set of words. Until then, stem-cell biologists should stick to debating the merits and ethics of their work using clear and simple language. They have a strong case to make that will not be helped by playing semantic games in an effort to evade scrutiny. ■

RESEARCH HIGHLIGHTS

Light as a feather

Phys. Rev. E (in the press)

The drab brown of male peacocks' tail feathers has been shown to arise from the same kind of structures that produce the vibrant colours of the 'eye'.

Photonic crystals give colour to many insects, butterflies and birds. The crystals have a periodic structure, often created by repeating patterns of tiny holes, that usually only reflects a narrow range of wavelengths of light.

Researchers from Fudan University in Shanghai, headed by Xiaohan Liu and Jian Zi, have now proved that a subtle arrangement of melanin rods and air spaces can also reflect the mix of colours needed to produce brown. The structure may inspire designs for artificial photonic crystals.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

D. A. MURAWSKI/GETTY IMAGES

DRUG DISCOVERY

Screen for promiscuity

Nature Chem. Biol. doi:10.1038/nchembio718 (2005)

Two new drug-screening techniques could reduce the amount of time that pharmaceutical companies waste exploring the biological properties of 'promiscuous inhibitors'.

Such molecules seem promising as potential drugs when tested in a high-throughput fashion, as they inhibit the activity of a target enzyme. But they often have this effect by forming aggregates that sequester the enzymes, rather than binding to the enzyme's active site. So promiscuous inhibitors can rarely be developed into drugs.

Researchers led by Brian Shoichet and Kip Guy, both from the University of California, San Francisco, have developed high-throughput screens that use detergents (to disrupt aggregates) and dynamic light scattering (to detect aggregates). When applied to 1,030 drug-like molecules, both techniques identified promiscuous inhibitors.

QUANTUM PHYSICS

Time's up

Phys. Rev. Lett. **94**, 230401 (2005)

There is a fundamental limit to how long quantum coherence can last, say Jan Zaenen and his colleagues at the University of Leiden, the Netherlands.

Coherence, which allows many particles to share the same quantum state, underpins phenomena ranging from superfluidity to quantum teleportation. It is also key to

proposals for quantum computers. But Zaenen's team shows that spontaneous fluctuations destroy quantum coherence in a time period that depends on the size and temperature of the system. For macroscopic bodies this can take centuries, but at the 'mesoscale' of hundreds of nanometres, it can happen in seconds. Fortunately, proposals for quantum computers don't tend to invoke mesoscale bits, so they are not undermined.

MOLECULAR BIOLOGY

Enlightened proteins

Chem. Biol. **12**, 685–693 (2005)

Light can act as a powerful switch to set off chemical reactions. Harnessing this effect, researchers led by Timothy Dore of the University of Georgia in Athens, and Erin Schuman of the California Institute of Technology, Pasadena, have developed a system that could be used to inhibit protein synthesis in cells in a controlled way.

They bound an antibiotic compound that hampers protein formation to a light-sensitive molecule called a chromophore. The antibiotic is released, and thereby activated, when the compound is exposed to ultraviolet light.

STRUCTURAL BIOLOGY

All in the ions

Cell **121**, 1005–1016 (2005)

The enzyme RNaseH is a member of the nuclease family of enzymes that cut strands of DNA and RNA. In work that may help to explain how the various nucleases target

different substrates, researchers led by Wei Yang at the National Institutes of Health, Bethesda, have described the structure of RNaseH bound to a DNA–RNA hybrid.

RNaseH acts on molecules that contain both DNA and RNA, but cuts only the RNA chain. The structure shows that two metal ions are involved in catalysing cleavage of the RNA strand. Nucleases related to RNaseH also use two metal ions to make cuts, but the geometric arrangement of the ions in these nucleases is different.

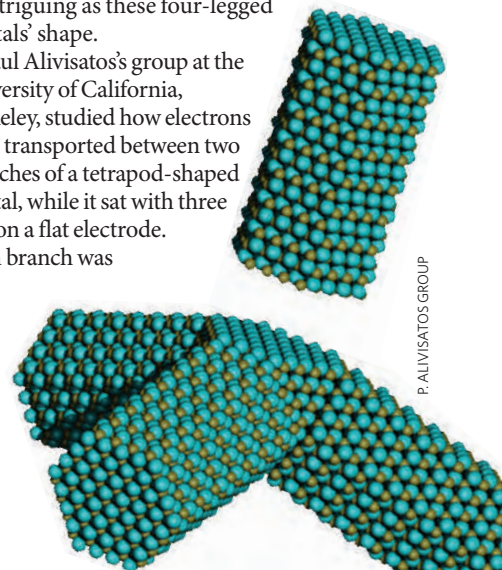
SOLID-STATE PHYSICS

Bendy legs

Nano Lett. doi:10.1021/nl051064g (2005)

The first investigation of the electrical properties of semiconductor tetrapods (pictured), has revealed a behaviour that is as intriguing as these four-legged crystals' shape.

Paul Alivisatos's group at the University of California, Berkeley, studied how electrons were transported between two branches of a tetrapod-shaped crystal, while it sat with three legs on a flat electrode. Each branch was



P. ALIVISATOS GROUP

150 nanometres long. Depending on how bent the branches had become when the tetrapod stuck to the electrode surface, electrons either behaved in a quantum way and hopped across the central junction, or a simple current flowed. This unexpected complexity could be put to use in nanoscale circuits, where the tetrapod's branches act as interconnects.

IMMUNOLOGY

On the defensive

J. Clin. Invest. **115**, 1806–1815 (2005)

Tissues attacked by bacteria can defend themselves by increasing levels of certain molecules, such as nitric oxide, that kill microbes. A protein called hypoxia-inducible factor 1 is already known to control this response in tissues that lack oxygen — a typical sign that they are under attack. Now, increased production of this protein by white blood cells has been directly linked to the presence of bacteria by Randall Johnson and Victor Nizet of the University of California, San Diego, and their colleagues.

Johnson's team also showed that mice were more susceptible to infection if they lacked the protein. But the researchers' suggestion that therapies that enhance production of the protein could boost a patient's immune system, although shown *in vitro*, remains to be tested *in vivo*.

DEVELOPMENTAL BIOLOGY

Uncertain destiny

J. Exp. Med. doi:10.1084/jem.20050146 (2005)

A single type of precursor cell found in the thymus can develop into a T cell, B cell or dendritic cell, a study finds.

The precursor cells migrate from the bone marrow to the thymus before maturing into these immune cells. To pinpoint the moment when the cells' fates are sealed, Claudia Benz and Conrad Bleul from Freiburg's Max-Planck Institute for Immunology monitored the expression of a receptor protein that marks an early stage of T-cell development. The researchers identified a threshold of expression in the most immature precursors that marks an important branching point in the cell hierarchy. Beyond this threshold, the cells were incapable of turning into B cells.

CANCER

Liver trouble

Cell **121**, 977–990 (2005)

One link between inflammation and cancer is known to involve the NF- κ B pathway, which regulates gene expression. Now Michael Karin and his colleagues at the

University of California, have implicated the pathway's activator, IKK β , in chemically induced liver cancer.

Mice injected with a chemical carcinogen were more prone to cancer when IKK β was deleted from their hepatocyte liver cells. But if the enzyme was deleted from both hepatocytes and Kupffer cells — a type of immune cell in the liver — the animals were less likely to develop cancer. Karin's lab conclude that carcinogenesis depends on 'crosstalk' between damaged hepatocytes and Kupffer cells, mediated by IKK β .

BLICKWINKEL / ALAMY

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

ANIMAL BEHAVIOUR

Smart settlers

Proc. R. Soc. Lond. B doi:10.1098/rspb.2005.3099 (2005)

Birds with bigger brains are less likely to fly south for the winter, a survey of the habits of 134 species has shown. This supports the theory that migration evolved in birds that weren't smart enough to survive cold weather.

Daniel Sol from the Autonomous University of Barcelona and his colleagues analysed existing data on birds living in temperate regions of Europe, Scandinavia and western Russia. In addition to finding that non-migratory birds have larger brains relative to their body size than species that migrate, they also discovered that non-migratory birds, such as the blackbird (*Turdus merula*, pictured), are more flexible in their feeding habits.

JOURNAL CLUB

Tony Pawson

Samuel Lunenfeld Research
Institute, Mount Sinai Hospital,
Toronto, Canada

**Signalling proteins have taught
this week's writer that cells can
add one and one to get three.**

The thought that one might understand mechanistically how cells work has always struck me as heady stuff. Yet, although we know a great deal about the pathways that orchestrate cells' responses to internal and external signals, and about regulatory systems such as the cell cycle, more complex cellular functions largely await explanation.

One exciting possibility is that cells are displaying emergent properties. My own research is focused on signalling proteins, which govern the dynamic behaviour of cells, and I am enthusiastic about the notion that signalling proteins might interact to yield an activity that is more than the sum of its parts.

Recently, Kevan Shokat and colleagues demonstrated such emergent behaviour for two protein kinase enzymes in the yeast *Saccharomyces cerevisiae* (C. Kung *et al. Proc. Natl Acad. Sci. USA* **102**, 3587–3592; 2005). By analysing gene expression, they examined the effects of inhibiting either kinase alone, or both together.

They found that one set of genes was affected by inhibition of the Cdk1 enzyme, which controls aspects of the cell cycle, and another was regulated by Pho85, which is involved in phosphate metabolism. The key experiments showed that simultaneously inhibiting both kinases also regulated a third set of genes, which participate in cell budding.

These findings are clinically important because the inhibitors used to treat diseases such as cancer tend to affect multiple kinases. Such non-specificity was initially viewed as a failing, but Shokat's data suggest that inactivating more than one kinase may be advantageous, or even critical, for a therapeutic effect.

NEWS

Flu officials pull back from raising global alert level

The world last week seemed to edge closer to the brink of a flu pandemic. On 30 June, officials at the World Health Organization (WHO) revealed that they recently considered raising the threat level of a global pandemic, from the current 3 on a six-point scale, to 4 or even 5.

The scare was triggered a few weeks ago when several research groups visiting Vietnam filed preliminary reports that many people with mild cases of influenza — and those in contact with them — were testing positive for the deadly avian flu strain H5N1. This suggested that there was widespread human-to-human transmission of the virus.

Subsequent tests have so far failed to confirm this, and WHO spokesman Dick Thompson is keen to play down the incident. “It was just unpublished information provided to us in preliminary form that spurred an investigation,” he says. “We thought about upgrading the alert. We looked at it fast and strongly, and based on that decided not to upgrade.”

But take a closer look, and the picture in Vietnam is one of confusion rather than reassurance. The first signs of trouble came in May, with reports of small clusters of human cases of H5N1, including a rise in the infection of older people and an increase in milder cases — all signs consistent with the possibility that the virus had mutated to achieve improved, although still inefficient, human-to-human

spread (see *Nature* **435**, 391; 2005).

Concern mounted in subsequent weeks as several international groups investigated the clusters using different methods, including the polymerase chain reaction (PCR), which amplifies DNA sequences, and western blots, which use antibodies to detect proteins. Despite using different tests, each of the teams reported that “substantial proportions” of the hundreds of people it had tested seemed to be infected with H5N1.

That led the WHO to consider upgrading the pandemic threat level to 4 (small, localized clusters of human infection) or 5 (large clusters of infection) — just one step away from a full-blown global pandemic. But first it asked an international team of experts, including Masato Tashiro, a virologist at the National Institute of Infectious Diseases in Tokyo, to retest many of the samples and some new ones, using the WHO’s own PCR tests.

They found no evidence of clusters of human-to-human transmission. “This is good news,” Tashiro says, relieved that his worst suspicions weren’t confirmed. But it remains unclear why the various groups got different results.

Samples have now been sent to a WHO laboratory in Hong Kong for the last word in confirmation: antibody neutralization assays. These take time as they involve growing large amounts of the virus for analysis, but a firm

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Close to danger? A young boy feeds ducks, a source of the H5N1 flu virus, in Vietnam.

conclusion is expected by the end of the month.

In the meantime, the WHO is holding off raising the alarm. “Because of the consequences of such a change, the WHO is following a cautious approach,” it said in a statement last week. Pushing the level to 4 for the first time would mean deploying the international stockpile of antiviral drugs to try to contain or stamp out the spread, and would probably result in countries restricting travel to Vietnam.

But Tashiro remains concerned that he and his colleagues didn’t have enough time to check

Infected birds poised to take flu virus south

Thousands of migratory birds infected with H5N1 avian flu virus at Qinghai Lake in western China pose a serious risk of spreading the disease to southeast Asia, India, Siberia, Australia and New Zealand when they fly home this September, scientists are warning.

The outbreak first hit the headlines in May (see *Nature* **435**, 542–543; 2005). Before then, deaths from H5N1 in migratory birds were limited, and many suspected they were dead-end hosts that occasionally picked up infection from poultry. But the virus seemed to have mutated into a more virulent form, and within weeks more than

6,000 birds of five different species had died at the breeding site.

Late last month, the Chinese government finally allowed 17 experts from the World Health Organization (WHO) to visit part of the quarantined area. Although their movements were restricted and they were unable to meet or interview any local people, the scientists were allowed to take virus samples for testing.

DNA sequences of those samples by scientists from Hong Kong, China and the United States are published this week^{1,2}. They confirm that the virus is a new form of H5N1, clearly

distinguishable from the strains seen in Vietnam and Thailand, and more closely related to variants isolated from poultry in China or Hong Kong. Experimental infection of mice and chickens also showed that the new variant is highly virulent² — comparable to the Vietnamese and Thai strains that have caused fatal human cases.

It is unclear how China will cope with such a huge outbreak. At a press conference in Beijing on 28 June, WHO officials complained that China had tested only a handful of birds from the five affected species, and none at all from the other 184 species present

at the lake, which could be acting as carriers. More than 100,000 birds will fly home from the lake in the next few months, and the officials urged China to implement large-scale testing along with a control strategy, possibly involving mass culling or vaccination of wild birds, before September. Experts are also still waiting for clearance from China to visit three subsequent outbreaks of the H5N1 virus in migratory birds in Xinjiang province, west of Qinghai. **D.B.**

1. Chen, H. *et al.* *Nature* doi:10.1038/nature03974 (2005).
2. Liu, J. *et al.* *Science* doi:10.1126/science.1115273 (2005).

**G8 SUMMIT**

See *Nature's* coverage of this week's meeting of the eight industrialized nations online.

www.nature.com/news

HOANG DINH NAM/AFP/GETTY IMAGES

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

all of the clinical and epidemiological information associated with the initial lab samples. Follow-up work is also complicated by the fact that recovered patients have now returned home, making it hard to track down people they might have infected. "We still have a big problem collecting enough good data," he says.

Jeremy Farrar, director of the Wellcome Trust Clinical Research Unit at the Hospital for Tropical Diseases in Ho Chi Minh City, says that much of the uncertainty over the prevalence of H5N1 could be avoided if Vietnam had better facilities for testing samples locally. "The international community continues to suggest that countries ship samples out somewhere else," he says, "while doing absolutely nothing to invest in enhancing the scientific capacity of the Vietnamese to respond to the epidemics themselves."

In the meantime, Tashiro adds that even if final tests confirm his negative results, "the fundamental situation has not changed". Many are concerned that July and August will bring a new and bigger wave of flu cases in Vietnam, as happened last year during the hot rainy season. And recent events in China bode just as ill. Scientists investigating migratory birds infected with H5N1 in western China are now warning that these pose an explosive risk of spreading the virus along their migration routes as they fly south in September (see 'Infected birds poised to take flu virus south', opposite).

At a UN meeting on bird flu held this week in Kuala Lumpur, Malaysia, the WHO's western Pacific regional director, Shigeru Omi, warned that H5N1 remains at a "tipping point". ■

Declan Butler

Climate change: is the US Congress bullying experts?

Last week, US Congressman Joe Barton, head of the House of Representatives' energy and commerce committee, wrote to three leading climate researchers, the head of the National Science Foundation and **Rajendra Pachauri**, chairman of the Intergovernmental Panel on Climate Change (IPCC). Barton asked for extensive information about their careers, funding and research.

The letters focused on a 1998 finding by Michael Mann of the University of Virginia, Charlottesville, that has been dubbed the 'hockey stick' graph because of its shape. Mann's study created headlines around the world when it suggested that the twentieth century was the warmest of the past millennium, and the 1990s was the warmest decade. The finding is central to the IPCC's most recent assessment of climate change.

Scientists have called the aggressive tone of the letters disturbing and dangerous (see page 1). They have accused Barton of attempting to bully climate researchers with whom he does not agree. *Nature* asked Pachauri for his reaction, and found him undaunted.

What was your first thought when you read the letter?

I was very surprised. This is the first time I have received a letter of this nature.

Do you feel obliged to respond?

I will first consult my colleagues in the IPCC. Over the next days we will decide whether and how to react. We might not do anything at all.

What kind of information would you consider providing?

I would not hesitate, out of courtesy, to provide basic information about how the IPCC functions and about the manner in which we choose our authors. This is a well established and absolutely transparent process. The only criteria are scientific merit and integrity. I don't think we need to provide more information than that. I guess it will be sufficient to mention the processes and procedures of the IPCC and to refer the committee to our website.

N. PISARENKO/AP

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Climate chief Rajendra Pachauri was surprised to receive a letter from a US House committee.

Is it appropriate for a US House committee to make these demands?

Yes, we're living in a democracy. But I don't know how anyone outside the scientific community would be able to make use of the information — it would take weeks or months to process all the information that is requested.

Was it unwise to give Mann's 'hockey stick' so much prominence in the IPCC's summary for policy-makers?

No. It is no exaggeration and it doesn't contradict the rest of the IPCC assessment. Of course you can always argue about details. But we assess all the available literature, and we found the hockey stick was consistent with that.

Do you think individual scientists such as Mann need to be better protected against pressure from politicians?

The IPCC cannot do that. But Mann and his colleagues are distinguished, independent scientists who are able to explain their points of view. These letters don't curb their independence. And the recipients don't need to provide all the information requested. By and large, I don't regard this as a threat to the scientific community. ■

Interview by Quirin Schiermeier

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Parasite infiltrates fruitfly research

LONDON

An unnoticed parasite could be scuppering fly geneticists' experiments. Up to a third of laboratory strains of their favourite test organism — the fruitfly *Drosophila melanogaster* — are infected with a parasite that affects the flies' biology, according to a study of hundreds of stocks of the fly. The finding suggests that the bacterium could be silently affecting the outcome of a significant number of studies.

The parasite, *Wolbachia*, can only be transmitted through infected eggs, and has evolved strategies for maximizing its transmission, such as switching the sex of male embryos. It can reduce *Drosophila* sperm and egg production, influence longevity and — critically for experiments — shelter its host from the effects of harmful mutations.

One mutation in a gene called *Sex-lethal* prevents female flies from producing eggs, for example, but *Wolbachia*-infected females carrying this mutation produce near-normal numbers of eggs¹. Certain strains with differing lifespans live for the same length of time after being treated with the antibiotic tetracycline, which kills *Wolbachia*².

Yet *Drosophila* researchers rarely consider *Wolbachia*'s effects, according to Timothy Karr of the University of Bath, UK. "It's amazing how many people don't know this," he says.

Karr and his team surveyed 609 of the

mutant strains housed at the Bloomington Stock Center at Indiana University for *Wolbachia* genes. They picked a wide range that included strains with large DNA deletions called deficiencies, smaller 'point' mutations, and mutations created with mobile pieces of DNA called P-elements, as well as normal, or 'wild-type', stocks. Some 30% of strains were infected, they reported³.

The team also discovered an unidentified mutation that kills flies when combined with another mutation called *chico*. But with the parasite present, such flies survive. "Now you have a new mutant that should have killed the fly," says Karr. "*Wolbachia* is keeping it alive."

Finally, they studied the distribution of the bacteria in fly larvae, and found that every tissue was infected. "It raises the bar of the alarm," says Karr. In theory this could affect any characteristic, he warns.

It looks as though no *Drosophila* geneticist can afford to ignore the possibility of *Wolbachia* influencing their results. "Now the paper is coming out, people are going to have to deal with it," says Karr. "It remains an open question how much of an effect there will be."

Other researchers say Karr's findings are important, but *Drosophila* genetics is not in any kind of crisis. "These are complicated animals," says Gerry Rubin, a *Drosophila* genet-

cist based at the Howard Hughes Medical Institute near Washington DC. The idea that an environmental factor could be affecting the flies' biology is nothing new, he says. "It isn't the first thing we didn't know about and it won't be the last." Knowing about these effects means researchers can now take them into account in experiments. "You can't control for things you don't know about," Rubin says.

Kathy Matthews, co-director of the Bloomington collection, says there are no plans to 'cure' the infected stocks with antibiotics. "This paper will help alert *Drosophila* researchers to the possibility of the presence of *Wolbachia* in any given stock, which is certainly a good thing," she told *Nature*. "But we view it more as a useful point of information about *Drosophila* biology than a call to arms."

That's good news for Karr, who sees the infected strains as a tremendous resource for probing the interaction between *Wolbachia* and its host, and perhaps for learning about how parasites make the transition to becoming symbionts. He also hopes to undertake a systematic study of all the mutants at Bloomington, comparing cured with uncured lines to see how they are affected by the parasite. ■

Claire Ainsworth

1. Starr, D. & Cline, T. *Nature* **418**, 76–79 (2002).
2. Driver, C., Georgiou, A. & Georgiou, G. *Biogerontology* **5**, 185–192 (2004).
3. Clark, M. E., Anderson, C., Cande, J. & Karr, T. L. *Genetics* doi:10.1534/genetics.104.038901 (2005).

Stem-cell 'heroes' celebrate a series of breakthroughs

SAN FRANCISCO

For years, dark clouds of controversy have obscured the horizons of stem-cell research. But at last month's annual meeting of the International Society for Stem Cell Research, scientists were buzzing with optimism.

"The tide is turning," said Len Zon, the society's outgoing president, at the meeting's opening session in San Francisco. "The field looks very bright."

Zon and the 2,100 other attendees have plenty to celebrate. One chamber of the US Congress has voted to ease restrictions on federal funding for stem-cell research, and the Senate may take up the issue this month. The creation of state stem-cell research initiatives — including the \$3-billion California Institute for Regenerative Medicine (CIRM) — has radically altered the research landscape in the United States.

And the recent report from South Korean researcher Woo Suk Hwang of Seoul National University of the first embryonic stem-cell lines tailored to individual patients (see *Nature* **435**, 393; 2005) has pushed the field closer to clinical reality. Stem-cell scientists are championing at the bit to visit Hwang's lab and study his techniques. "There are a lot of people in line to go," says delegate Renee Reijo-Pera of the University of California, San Francisco.

There was plenty at the meeting itself to keep researchers buoyant. Kevin Eggan of the Harvard Stem Cell Institute in Boston reported that his lab has fused a human embryonic stem cell

with an adult skin cell and that the stem cell 'reprogrammed' the skin cell's nucleus, causing it to act like an embryonic stem cell. He still has to work out how to remove the stem-cell DNA from the fused cell. But if he can, the technique could be a way to produce new stem-cell lines matched to the DNA of a patient, without having to use a donor egg or create an embryo — the two sources of most of the ethical concerns surrounding embryonic stem-cell research.

Meanwhile, pioneering stem-cell biologist James Thomson of the University of Wisconsin-Madison reported progress in making stem-cell lines free of contamination by contact with material from other animals. This is crucial for making lines suitable for use as treatments in people. All cell lines currently available under US federal funding restrictions are contaminated in this way, so Thomson says that he hopes the research will help push the United States to change its rules.

Some investigators cautioned patience. Lorraine Young of the University of Nottingham, UK, unveiled the first complete profile of epigenetic modifications in human stem cells, and warned that cultured cell lines can change rapidly over time. This needs to be better understood before pressing ahead with clinical research, she said.

But CIRM chairman Robert Klein felt there was no time to lose. "You are my heroes," he told the meeting. "We must work quickly, or the forces against us may reverse the tide." ■

Erika Check

"We must work quickly, or the forces against us may reverse the tide."

Clear thinking: Robert Klein, chairman of the California Institute for Regenerative Medicine, puts his mind to the challenges ahead.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

ON THE RECORD

“We used to be afraid of comets. The dinosaurs should have been afraid of comets. Now it is the comet’s turn to be afraid.”

UCL space researcher Andrew Coates gets a little carried away describing NASA’s Deep Impact mission, in which a probe crashed into the Tempel 1 comet (see page 13).

“The guidance offered in this article on how to anticipate, model and minimize a botulinum toxin attack can be valuable for biodefence.”

Bruce Alberts, president of the US National Academy of Sciences, justifies publishing a paper modelling a bioterrorist attack with botulinum toxin.

“If something bad happens as a result of this, it’s the Department of Health and Human Services who will have to deal with it, not the academy.”

US health department spokesman Marc Wolfson explains the agency’s objections to the botulinum paper.

SCORECARD

**Poisson distribution**

Thai fishermen have landed a whopper: a giant catfish tipping the scales at 293 kilograms. Making a splash as the largest freshwater fish on record, conservationists were keen for the behemoth to be set free. Sadly, it had already cashed its chips — cooked and eaten by its proud captors.

**Golden touch**

Isaac Newton’s long-lost notes on alchemy have been found — but scholars are struggling to turn the written code into something they can read.

**Hard to swallow**

A study finds that processed duck meat exported from China to Japan in 2003 contained bird flu virus, giving rise to fears that it could be a threat to human health.

Scientists finally get their hands on Kennewick man

SAN DIEGO

America’s most highly contested anthropology specimen, Kennewick man, is finally being studied by scientists.

After nine years of federal-court battles and several months of preparation, researchers last month began examining the ancient skull and bones. The US government and Native American tribes had fought to block the bones’ examination under a federal law designed to protect ancient human remains. But last year, the eight scientists involved won the legal battle.

The male skeleton was found in July 1996, along the banks of the Columbia River near Kennewick, Washington. Preliminary radiocarbon dating suggests that Kennewick man lived about 9,300 calendar years ago. As there are so few full skeletons of New World individuals more than 8,000 years old, the researchers want to closely catalogue everything about the skeletal remains to try to work out who he was and where he came from. The bones will be inventoried and measured, and the skeleton reconstructed. It will then be checked for evidence of disease, trauma, diet and, with luck, DNA. The shape of the skull and the length of the arms and legs are also particularly valuable for deducing an individual’s evolutionary history.

So far, key skeletal remains — particularly pieces of the skull — have been scanned using computerized tomography. The researchers aim to make a complete skull cast from the

scan data to enable them to study the skull without disturbing the real bones too much. A stone blade or point encased in the pelvis has also been extensively scanned. The weapon is buried so deeply that the researchers can’t identify its point or base, which could help to determine the weapon’s heritage.

Douglas Owsley, head anthropologist at the Smithsonian Institution in Washington DC, who is leading some of the studies, says he is

glad the work has finally begun.

“I am really relieved,” he says.

“This is one of the most important specimens in the Americas.”

But not all of the scientists who fought the case can take part in the work. The battle in federal court in

Portland, Oregon, was so long that two of the eight scientists have retired, and a third, Robson Bonnicksen of Texas A&M University in College Station, died in his sleep last December, aged 64 — just two weeks after getting a look at the specimen. “He was euphoric when he got to see it,” says Cleone Hawkinson, a retired anthropologist from Portland who helps the group. Ironically, Bonnicksen had filed a court declaration in 2002 arguing for a quick decision because senior scientists in the group might not live to study the bones.

The dispute has also been financially costly. The federal government must pay the scientists’ \$2.5-million legal costs, and the government’s own costs are estimated at \$6 million. ■

Rex Dalton

“This is one of the most important specimens in the Americas.”

Within grasp: nine years after its discovery, the true origins of the 9,000-year-old skeleton of Kennewick man can at last be investigated.



H. E. BERRYMAN/LEBANON, TENNESSEE



**OCEAN LIFE IN TROUBLE
AS ACID LEVELS RISE**
Report calls for more
stringent carbon cuts
to protect seas.
www.nature.com/news

S. GSCHMEISSNER/SPL

NASA

US willing to pay for Russia's help in space

WASHINGTON DC

The Bush administration finally moved last week to loosen the legal ties preventing NASA from buying Russian transportation services for the International Space Station.

The 2000 Iran Nonproliferation Act (INA) bars the US space agency from directly purchasing Russian goods or services, including astronaut rides on the Soyuz vehicle (pictured) or cargo deliveries on the Progress capsule. As a principal partner in the space station, Russia has been providing both services free of charge since the station was first occupied in 2000. But that agreement will expire in April 2006. Without access to the Soyuz as an emergency 'lifeboat', NASA could only have astronauts on the station when the space shuttle was docked. As NASA administrator Michael Griffin told the House Committee on Science last week, "the United States cannot effectively utilize the space station without our Russian partners".

The INA was intended to put pressure on Russia not to sell missile and weapons technology to Iran. But many experts believe it has been counterproductive. A report on US-Russian space cooperation published by the Washington-based Eisenhower Institute earlier this year concluded that it has hampered positive engagement with Russian space scientists and engineers, and become a

"blanket sanction which affected not only the proliferators but also those complying with nonproliferation standards".

And as NASA's plans for its own space-station rescue vehicle have faltered, the law has become a serious obstacle for the space station. "This was mostly damaging the United States," says University of Maryland physicist and former Russian space official Roald Sagdeev.

Failed strategy

Last week, Griffin and Secretary of State Condoleezza Rice wrote to the House science committee chairman Sherwood Boehlert saying they want to amend the INA — presumably to allow NASA to pay for future Soyuz and Progress flights. The specifics of the amendment, which would have to be passed by Congress, are expected as early as this week.

Sagdeev credits Griffin with solving a dilemma that his predecessor, Sean O'Keefe, had avoided facing. "Finally Mike Griffin did the right thing," he says.

Even science committee member Dana Rohrabacher (Republican, California), a staunch cold warrior who has been one of the principal defenders of the INA, was ready to admit defeat last week.

"It was a worthy effort at the time to make sure that we pressured the Russians not to

participate in the developing of a nuclear facility in Iran," he told a science committee hearing last week. "That strategy has, however, not worked."

The proposed White House amendment will presumably leave most of the INA intact while making an exception for space-station transportation services — an approach that Sagdeev calls "the victory of common sense over ideology".

That would solve an embarrassing problem for NASA, but will not necessarily signal closer ties between the United States and Russia in space. Indeed, Griffin sent mixed messages on that topic last week. "Among the best things to have come from our space programme over the last 15 years is the space cooperation that we've enjoyed with Russia," he told the committee. Later, however, he listed Russia as one of NASA's main competitors in space, and said, "My goal is to see to it that America is always in the lead in that competition. That matters greatly to me."

Sagdeev says Griffin's view of Russia as a competitor is misguided, given that country's "miserable space budget". But he adds that cooperation with Europe is now a higher priority for the Russian space programme than cooperation with the United States.

Tony Reichhardt

Genetic patent singles out Jewish women

MUNICH

Europe's rules for gene patenting are under attack for allowing racial discrimination. On 29 June, the European Patent Office (EPO) upheld a patent that will mean Ashkenazi Jews have to pay for screening for a particular breast-cancer gene.

The patent, on the *BRCA2* gene, was filed by Myriad Genetics of Salt Lake City in Utah, although the company has since transferred ownership to a consortium that includes the University of Utah Research Foundation. It was submitted in 1996, granted in 2003, and originally covered all possible research and diagnostic tools involving the gene. After challenges, the consortium narrowed the patent's scope to cover the diagnosis of a particular mutation in Ashkenazi Jews.

In January, a European patent that Myriad filed on the related *BRCA1* gene was revoked, because critics revealed errors in the DNA

sequence as first registered (see *Nature* 433, 344; 2005). The *BRCA2* sequence seems to be correct, but the consortium has refocused its *BRCA2* patent on the point where its claims of intellectual ownership are strongest — a mutation that seems to be found exclusively in Ashkenazi Jews.

The EPO decision means that European patients will be asked whether they are Ashkenazi. If they are, they will have to pay the same high royalties for *BRCA2* testing that US doctors have to demand from all patients — the US patent office has granted broad versions of both the *BRCA1* and *BRCA2* patents.

The Munich-based EPO has granted patents on hundreds of individual genes. But very few patent holders demand licence fees from public health clinics. When Myriad bucked that trend, many European clinics decided to ignore its royalty requests. And various geneticists challenged Myriad's patents.

**"This is a moral
not a legal issue."**

Those challengers are unhappy with last week's result. "We are thinking about whether to appeal again, but given the general rules of the EPO, the basis on which we might be successful is not obvious," says Gert Matthijs of the Centre for Human Genetics at Belgium's Catholic University of Leuven.

Dorit Lev, head of the Israel Association of Medical Geneticists, says that the situation is unacceptable. Ashkenazi Jewish women with this mutation frequently develop breast cancer. A simple genetic test identifies those at risk and screening is routine. "It's not right that they should be discriminated against," says Lev.

The only option for opponents is to press for changes to the grounds on which the EPO grants patents. "This is a moral not a legal issue and we are now thinking about challenging the political basis of the European rules," says Mary Rice, spokeswoman for the Vienna-based European Society of Human Genetics. ■

Alison Abbott

Lawmakers slash funds for nuclear research facility

The US Senate voted on 1 July to cut all construction funding for a major nuclear-weapons research facility in California, despite the fact that the government has already spent \$2.8 billion on it.

The National Ignition Facility (NIF) at the Lawrence Livermore National Laboratory in California is designed to use 192 lasers to study fusion and investigate atomic reactions such as those that take place inside nuclear bombs. The complex is 80% complete.

But the Senate sliced \$146 million that was earmarked for NIF from a Department of Energy funding bill for the 2006 fiscal year. Senator Pete Domenici (Republican, New Mexico) led the charge to stop funding for NIF. Facilities in his state compete for limited funds from the Department of Energy.

Officials at Lawrence Livermore say they are very concerned about the survival of NIF, which now employs about 1,000 scientific staff and has faced sharp scrutiny before for running over budget (see *Nature* 401, 195; 1999).

Retract emissions pledge, exhorts Russian sceptic

A prominent Russian scientist, well known for his scepticism about global warming (see *Nature* 431, 12–13; 2004), says that the Russian Academy of Sciences was in error when it signed a joint statement on climate change along with ten other national science academies.

Last week, Yuri Izrael, chairman of a climate committee for the Russian academy, said that the group does not in fact support the 8 June statement, which calls for immediate action to slow the emission of greenhouse gases. He called on the academy's president, Yuri Osipov, to reconsider his signature.

The US National Academy of Sciences has also run into a few stumbling blocks over its participation. It endorsed the statement but disavowed a strongly worded press release about the agreement, issued by the London-based Royal Society — another signatory.



NASA/JPL-CALTECH/UMD/GETTY IMAGES

Crunch point: NASA probe wallops comet

In a perfectly targeted dive, part of NASA's Deep Impact spacecraft smashed into comet Tempel 1 at 5:52 GMT on 4 July, releasing a spray of ancient debris from the comet's core.

The first pictures show a much larger plume than expected (above), suggesting that the comet's crust is more fragile than was thought. The debris contains material left over from the formation of our Solar System; further analysis

should reveal a cocktail of chemicals that once bombarded the primitive Earth.

NASA pushed ahead with the mission despite a US\$300-million lawsuit brought by Russian astrologer Marina Bai, who claimed that the collision could damage the natural balance of the Universe.

For an in-depth account of the impact, see www.nature.com/news.

Pesticide testing breeds discontent in the Senate

The US Senate has voted to ban the Environmental Protection Agency from testing pesticides on humans or using data from any previous such tests.

The ban was put in place during Bill Clinton's administration, but has recently lapsed. Senator Barbara Boxer (Democrat, California) reintroduced it, criticizing several ongoing pesticide studies.

The Senate also passed a competing amendment, which requires the agency to review the conditions in such tests but does not ban them outright. Negotiators will have to find a compromise between the two.

Bush bumps up aid for malaria control in Africa

The US president, George W. Bush, has said he will send \$1.2 billion in aid to Africa over the next five years to fight malaria.

The investment, announced last week in advance of the G8 summit, will target 175 million people in at least 15 countries.

Tanzania, Uganda and Angola will receive money next year; others will benefit later.

Public-health experts welcomed the promise of more aid, and hope the money will pan out as planned. Malaria kills more than a million people annually — most of them are children in Africa.

California legislators fail to hobble stem-cell research

California's \$3-billion stem-cell programme has survived a legislative attempt to make its funding and research more transparent, clearing the way for the first allocation of funds later this year.

Certain state legislators argued that there should be a statewide ballot this November to address transparency in the programme (see *Nature* 435, 544; 2005). But the lawmakers failed to win support for their idea.

Leaders of the California Institute for Regenerative Medicine, which is gearing up to spend \$300 million a year on stem-cell research for the next decade, are drawing up disclosure policies to address some of the concerns.

Is this any way to save a species?

Thanks to the influence of a powerful US senator, more than \$120 million has been pumped into research on Alaska's endangered Steller sea lions in just four years. Rex Dalton asks what we've learned.

A hardy creature of Alaska's forbidding Aleutian Islands, the mighty Steller sea lion (*Eumetopias jubatus*) survives where few other mammals can. In freezing temperatures and battered by Arctic storms, bulls of the species can reach 1,000 kilograms.

But a scientific initiative to determine why the Aleutian population of Stellers has plunged over the past three decades is making them a wilderness legend for another reason. For some biologists, the Steller has become a symbol of how not to conduct complex ecosystems research. "The dollars were directed toward dealing with a political conflict," claims Larry Crowder, a population ecologist at Duke University in Durham, North Carolina.

Over the past four years, the US government has poured more than \$120 million into Steller research — a sum described by one biologist working on the species as "obscene". It dwarfs the funding for research on other endangered marine animals, several of which are much closer to extinction. Yet this highly politicized programme has so far failed to resolve a key question — whether fishing is responsible for the Steller's decline.

Steller sea lions, named after the German naturalist Georg Steller, who described them in 1741, were abundant in the northern Pacific until the 1960s. But by 1990, numbers off western Alaska had dropped so dramatically that the Steller was declared a 'threatened' species under the US Endangered Species Act. In western Alaskan waters, this listing did nothing

to halt the decline. So in 1997, federal officials upgraded the status of the population west of 144° W to 'endangered'. Today, the western population of Stellers stands at around 35,000 adults.

But the cause of the precipitous decline remains a mystery — particularly when the separate Steller population along Alaska's southeast coast continued to thrive. Some biologists blame disease, pollution, predation by killer whales, entanglement in fishing gear, or a combination of these. Others suggest that the cause is depletion of the western Stellers' prey, either as a result of local climate change affecting fish populations or because of commercial fishing.

Knowing why the population declined will help devise a plan to ensure its long-term survival, and the Steller Sea Lion Research

Initiative, launched in 2001, is supposed to provide some answers. But the idea of running such a huge research programme wasn't dreamed up by conservation biologists. Rather, it was the product of congressional intervention in a battle between environmental groups and fishing interests.

Environmentalists believe that the Stellers' decline is directly linked to commercial fishing — in particular trawling for walleye pollock (*Theragra chalcogramma*), a cod-like species whose roe is in high demand for sushi. In the late 1990s, environmentalists were getting a sympathetic hearing from the administration of President Bill Clinton, and were using the courts to try to get tighter controls on the fishing industry.

In 1998, a coalition of environmental groups filed a legal complaint alleging that the US



A. TRITES/UNIV. BRITISH COLUMBIA



Million-dollar babies: endangered western Steller sea lions enjoy unprecedented research funding.

National Marine Fisheries Service (NMFS) had violated the Endangered Species Act by failing to adequately assess the impacts of fisheries on the Stellers. The ensuing court battle went through numerous twists and turns. But when the federal judge hearing the case blocked a trawling season in Alaskan waters in the summer of 2000, a powerful ally of the fishing industry stepped into the fray: Republican Senator Ted Stevens.

Alaska may not have many voters, but in Stevens it has one of the most senior and influential members of Congress. Then chairman of the Senate Appropriations Committee, Stevens could put his spin on almost any government funding.

With the fishing barons enraged over the court actions, Stevens used his influence to the full. In the waning December days of the 2000 congressional session, he held up the spending bill for the entire US government for a week, until he got what he wanted. Behind closed doors, a deal was hashed out in which pollock fishing would continue, except in areas close to Steller rookeries and haul-outs. Meanwhile, the National Research Council (NRC) was asked to conduct an independent review of the evidence on the Stellers' decline. And the Steller Sea Lion Research Initiative, costing more than \$40 million in the first year alone, was launched.

"It was not a good way to create a research programme," says Andrew Rosenberg, now an ecologist at the University of New Hampshire in Durham, but formerly a government fisheries scientist involved in research on Stellers. Until 2001, the federal budget for Steller research hovered around \$1 million a year. Suddenly, a huge amount of cash was thrown at Steller studies, with little prior planning. "I thought we would be lucky if 10% of the money was well spent," says one government scientist, who asked not to be named. The

funds, administered by the NMFS, also had to be spent quickly within particular budget years, which limited the scope for long-term ecological projects.

Now scientists are taking stock of what has been achieved. Last autumn, at a four-day conference on sea-lion conservation and research in Anchorage, Alaska, experts heard many of the first presentations of projects from the Steller initiative. In the opening session, Tom Loughlin of the NMFS's National Marine Mammal Laboratory in Seattle, who formerly helped administer the initiative, asked provocatively: "What does \$120 million buy?"

Value for money?

Views on that point were divided. Among researchers funded through the initiative, there was great enthusiasm for projects that had provided basic knowledge of a difficult animal to study in a dangerous environment. There were presentations on diet, metabolism and physiology; on possible pollutants and pathogens; on reproductive performance and juvenile health; and on how far Stellers roam from their rookeries and haul-outs — important information in determining the size of no-fish zones around these sites to ensure adequate food for the sea lions. "These are very complicated issues," says Douglas DeMaster, director of the Alaska Fisheries Science Center in Seattle, which oversees the Steller research initiative. "That makes for big expensive studies."

But some scientists argue that important opportunities to understand the Stellers' population biology have been missed. "The studies mainly just allow us to understand Stellers better," says Crowder. "A study just can't be about metabolic pathways of a certain fish food. You have to link the study to population dynamics."

"I am pretty sceptical about the research money being targeted efficiently on key

issues," concludes Crowder, who served on the NRC panel that eventually reported in 2003, with the Steller research initiative already well down the tracks.

That report¹, which suggested that a combination of factors other than a diminished food supply was probably to blame for the Stellers' decline, was itself controversial. Some experts criticized the NRC panel for concentrating too much on the contested reasons for the Stellers' decline rather than advising on research and other measures that might be taken to ensure their survival. "It asked the wrong question", Rosenberg argues. "The council didn't jump into this the way it should have."

But the panel did come up with one firm recommendation for research: the best way to test the impact of commercial fishing, it said, would be to establish areas open and closed to fishing and monitor the survival of sea lions in each. In the event, getting the fishing industry to agree to this proved politically impossible. "We wanted to do it, but no one wanted an area closed for fishing," says DeMaster.

For environmentalists, this failure is indicative of the problem with the Steller research programme. They contend that the legislation was worded in such a way that money would be steered away from some of the most germane research — in particular, any projects that might implicate the fishing industry in the Stellers' decline. The legislation specified, for example, that the initiative should include projects looking at the roles of predators and climate change. "The money didn't come without strings," claims Janis Searles, an attorney in Portland, Oregon, who was formerly with Earthjustice, a non-profit public interest law firm based in Oakland, California. "From the US citizenry's perspective, having Congress direct the spending wasn't the smartest idea."

In the absence of any comparison between large areas open and closed for fishing, researchers have adopted other methods of investigating whether the Stellers are nutritionally stressed — and if so, why. The main bone of contention surrounds the role of pollock in the sea lions' diet, and an idea called the 'junk-food hypothesis'. This argues that Stellers prefer to eat oily fish such as herring and mackerel. But since the 1960s, the junk-food theory suggests, climate change has reduced the stocks of these fish in western Alaskan waters, causing them to be replaced by pollock, which is less nutritious².

Advocates of the junk-food hypothesis point to evidence that includes infrared imaging of Stellers preferentially feeding on herring at night in the wild³, and argue that pollock fishing is not the main reason for the animals' decline. The idea has been promoted most strongly by Andrew Trites, a biologist at the University of British Columbia in Vancouver, Canada — whose relationship with the fishing industry has sparked controversy (see 'Conflict of interest claims muddy the waters', overleaf).

At the Anchorage meeting, some researchers

L. FRITZ/NOAA FISHERIES SERVICE; SEALION RESEARCH PERMIT: 782-1532-00, 782-1532-02

Conflict of interest claims muddy the waters

In the lavishly funded world of Steller sea lion research, Andrew Trites is a stellar figure. At a major conference in Alaska last autumn, he co-authored 15 presentations on Steller sea lions — more than any other researcher. He is also research director of the North Pacific Universities Marine Mammal Research Consortium, which has received at least \$9 million from the Steller Sea Lion Research Initiative.

Yet some biologists are suspicious about the relationship between the fishing industry and Trites's consortium, which includes researchers at the University of British Columbia in Vancouver, Canada, where Trites is based, and scientists at three US universities: the University of Alaska, the University of Washington, and Oregon State University. It was set up in 1992, after fishing-industry leaders asked universities in the region to submit proposals for research on Steller sea lions.

In 1997, Trites wrote to US federal officials on behalf of the consortium, arguing that western Alaska's Stellers should not be classified as endangered. The move prompted Lloyd Lowry, then Alaska's top official managing marine mammal conservation, to resign from the consortium's scientific advisory board with



Researchers disagree over the impact of commercial fishing on the western Steller sea lions.

a stinging letter. Trites was "acting as [an] advocate for the commercial fishing industry instead of as [an] objective scientist," Lowry wrote. His letter accused Trites of misusing science in "a blatant attempt to absolve the groundfish trawl industry from any responsibility for the Steller problem".

With hindsight, Trites concedes that he should not have written on behalf of the consortium. He has also now changed his

mind about the Stellers' status, and agrees that the western population should be listed as endangered. And he rejects the accusation that he misused science to favour fishing interests. The consortium's research is focused on topics that its members believe are crucial to protecting Steller sea lions, he says. "We want to make sure we are asking the key questions to get the best value from the research."

claimed that the junk-food hypothesis doesn't stand up. Reviewing the available data on fish stocks and sea-lion diets, Lowell Fritz, a biologist at the Alaska Fisheries Science Center, argued that evidence for the shift proposed by Trites and his supporters is weak or non-existent. "It also is often overlooked that cousins of Stellers are doing just fine eating pollock all over the world," Fritz adds. His analysis will soon be published in *Marine Mammal Science*; Trites is planning to publish his own review supporting the theory.

While the debate rages on, one study presented at the Anchorage meeting did look squarely at the crucial relationship between fishing and Steller populations. Initiated before the congressional largesse and costing around \$200,000, it formed the doctoral thesis of Daniel Hennen, then a graduate student at Montana State University in Bozeman.

Hennen analysed population data for Stellers at 33 rookeries along the Aleutians from 1977 to 2001, relating the figures to regional fish-catch data for the same period. Before the animals were listed as threatened, he found that fishing seemed to be hurting the populations. After protection was in place, the populations studied started to bounce back.

"The decline in the Steller population was fastest in geographical areas where there was the most fishing," says Daniel Goodman, Hennen's supervisor at Montana State. Hennen, now a postdoc at the Alaska SeaLife Center in Seward, adds that the strict enforcement of a ban on sea-lion shooting after 1990 may also

have been significant, and that he plans to publish his results soon.

Crowder calls Hennen's work "the most strategic and carefully done research" on the relationship between fishing and the Steller population decline, noting an absence of similar studies in the published output of the Steller Sea Lion Research Initiative. "We are spending millions of dollars for grey literature that never has much influence," Crowder complains. "And this is happening at a time when

money for research on other endangered marine species — turtles, birds, otters and seals — isn't available."

More Steller research money is set to flow. President George W. Bush's budget for next year includes a proposal for about \$10 million, which could well be increased by Congress. A spokeswoman for Stevens says that the senator considers the programme valuable.

Meanwhile, federal officials, scientists and stakeholders, including fishing industry representatives, are still arguing over a new Steller recovery plan, to replace a version that dates from 1992.

And the political battles are destined to continue. In 2003, Stevens won another Congressional measure, designed to promote fishing in Alaskan waters. As a result, the state is now entertaining proposals to allow fishing in certain areas within 5.5 kilometres of the shore. "Some of these proposed fishing areas are near critical federally protected Steller habitat," says Michael LeVine, an attorney for Earthjustice based in Juneau, Alaska.

What does \$120 million buy? Not, it seems, an end to the controversy surrounding this majestic marine mammal.

Rex Dalton is Nature's US West Coast correspondent.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Senator Ted Stevens: the political dimension.

1. *The Decline of the Steller Sea Lion in Alaskan Waters: Untangling Food Webs and Fishing Nets* (Nat'l Acad. Press, Washington DC, 2003).
2. Rosen, D. A. S. & Trites, A. W. *Can. J. Zool.* **78**, 234–239 (2000).
3. Thomas, G. L. & Thorne, R. E. *Nature* **411**, 1013 (2001).

Deep in thought

Electrodes implanted in the brain could transform the lives of psychiatric patients. **Alison Abbott** watched an operation to release a man from his obsessive thoughts.



Neurosurgeon Volker Sturm threads a fine electrode into the brain of a patient.

“What are you thinking about?” Clamped to the operating table with his skull drilled through to expose a patch of brain, Herr Z. answers his psychiatrist anxiously: “I’m still thinking about the bad impression I must have made on the former colleague I met in the supermarket.”

It is close to midday and his brain surgery — which began at 8 am and will continue until 5 pm — has only partially distracted Herr Z. from this incessant, hammering thought.

Herr Z., now being operated on by Volker Sturm, a neurosurgeon at the University of Cologne, suffers from obsessive-compulsive disorder (OCD), which makes him unemployable. He rarely goes out, because he fears contamination from anything he touches. Twenty years of therapy have not helped.

An experimental surgical technique called deep brain stimulation (DBS) is currently the only hope for people like Herr Z. Having proved its value in the treatment of advanced Parkinson’s disease, neurosurgeons are now keen to try DBS to treat one of the world’s biggest health burdens — psychiatric disorders. Controlled clinical trials are under way in several centres in Europe and the United States to see if it can help with OCD and major depression. “The early results are very encouraging,” says Ali Rezai, a neurosurgeon at the Cleveland Clinic in Ohio.

DBS involves the insertion of electrodes deep in the brain to hit a precise neuro-

anatomical target that is believed to be central to the disease being treated. The electrodes are connected to a battery-driven stimulator that sends pulses of current to the target neurons and normalizes their activity. The stimulator is sewn into the belly or chest and can be switched on or off remotely through the skin.

The use of the technique has exploded since 1993, when Alim-Louis Benabid from the Grenoble University Hospital in France reported results from more than 80 patients with Parkinson’s disease¹. Around 30,000 similar operations have now been carried out worldwide. The success rate is high, and gratifyingly visible: a pulse of current through a correctly placed electrode instantly stops the tremors and releases the frozen muscles characteristic of the disease. The most common targets in Parkinson’s disease are the subthalamic nucleus and the globus pallidus, components of the brain’s motor circuitry whose signals are known to be distorted in Parkinson’s disease.

Deep targets

The motor circuit is one of several proposed neural pathways that circulate sensory information received from the outside world, by sight and touch for example, through the thalamus into the cortex, the ‘thinking’ part of the brain (see Graphic, opposite). Having ‘decided’ how best to respond, the cortex returns response signals through the thalamus: signals to move away from something, or to feel good

or bad about something, are shunted on to appropriate subcortical parts of the brain that arrange for the commands to be executed.

Surgeons wondered whether other, less well defined cortico-thalamic circuits could be targeted by DBS. The subcortical limbic system is especially interesting as it is associated with emotion and may be central to certain psychiatric conditions. Andres Lozano, a neurosurgeon at the University of Toronto who recently published results of the first trial of DBS in severe depression², makes no bones about it. “Our greatest opportunity is in the field of psychiatry.”

Surgeons are still not agreed about where best to strike with their electrodes. They realize that targeting one point will affect an entire circuit — and one modified circuit is likely to affect information flow through others. No one knows exactly how these various circuits, with their positive and negative feedback components, might weave through different anatomical structures.

But there are some pointers. In people who have suffered very localized brain damage, often following a stroke, behaviour can change in a very precise way. Neurosurgeons have often taken a lead for DBS from such observations and created carefully controlled damage, or lesions, in these areas to treat psychiatric diseases, including OCD. Bart Nuttin, a neurosurgeon from the Catholic University of Leuven in Belgium, who pioneered DBS in OCD patients³, targets a small frontal part of the

limbic system known as the internal capsule.

Other pointers come from functional magnetic resonance imaging (fMRI) and positron-emission tomography, which enable activity in different parts of the brain to be visualized. Lozano, for example, selected a limbic region known as Cg25 for stimulation after his collaborators found it to be particularly active in depressed patients.

Meanwhile, back on the operating table, Sturm is closing in on his preferred target, a particular section of the outer shell of the nucleus accumbens, a peanut-sized structure directly below the internal capsule.

Tracing a path

It is now 12:45 pm. Herr Z. began the day having his shaved head firmly clamped, under light anaesthetic, into a plastic stereotactic crown. The crown is now enabling Sturm to guide a 0.8-mm bore along the path he has defined on a computer screen with submillimetre precision. The path must hit the target exactly, without passing through any blood vessels — a haemorrhage could cause brain damage. But getting it right is hard as no two brains are exactly alike and brain-imaging techniques have limitations. After the clamping, Sturm took two different brain images: a computer tomography scan and a structural MRI scan. Sitting at the screen for more than an hour with a physicist and a neuroanatomist, he merged the images, identified his target and traced out the best straight line to its centre.

"I couldn't do this without working closely with basic scientists," says Sturm, now slowly threading a test microelectrode through Herr Z.'s brain along the set trajectory. The microelectrode converts neuronal activity to sound. Each anatomical structure has its own characteristic song, so this provides Sturm with a second guide. The recording is being carried out by a team of neuropsychologists, who are also taking the opportunity for an experiment of their own — to identify what may be characteristic neuronal firings when OCD patients try to shift their thoughts.

Sturm then inserts the true electrode. It is a hair-thin cylinder 6 mm long, with four 1-mm contacts spaced along it. Each contact can be individually manipulated from outside: its polarity, and the shape, frequency and voltage of the current, can be adjusted. If one contact doesn't have an effect, then the next, 1 mm away, may be in just the right position. An X-ray confirms that its position is spot-on.

1:30 pm. Time for the first test stimulations. Herr Z.'s psychiatrist, Michael Schormann, takes an envelope from his pocket and places it in Herr Z.'s hand. Envelopes upset Herr Z. — they might have been licked. His fingers grip the paper and he reports escalation to 10 on his personal anxiety scale. The test impulse does not reduce it. "It sometimes takes a while to get the effect, but it is psychologically important for patients to take an active part in the procedure," says Schormann. Relieved of

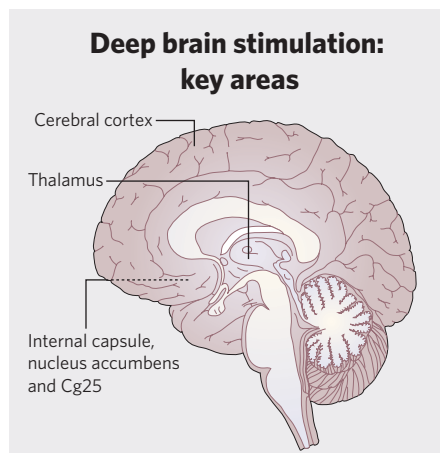


"I couldn't do this without working closely with basic scientists."
— Volker Sturm

the envelope, Herr Z.'s hand writhes above the sheet in a washing motion until Schormann wipes it clean with alcohol.

Herr Z. is wheeled out for an fMRI scan which confirms that his brain activity changes appropriately when the electrode is stimulated. Then back for the final sewing up. The long day is over. His stimulator and battery will be put in the next day.

Sturm is part of a small elite group of neurosurgeons in Europe and the United States,



including Nuttin, Lozano, Rezai and Benabid. Together, with their teams of basic scientists, they regularly meet to compare techniques and results in order to try to understand scientifically what they are doing.

This isn't simply a question of which anatomical target might be best, and why. The aim is also to understand better what the stimulation itself is doing. The original idea was that it simply blocked abnormal electrical activity, mimicking the effect of a lesion. Indeed, high-frequency pulses do damp down circuits by blocking excitatory neurons. But low-frequency pulses have the same effect by stimulating the inhibitory neurons in these circuits. The group also wants to improve three-dimensional neuroanatomy to help

make the positioning of electrodes more accurate and optimize their effects.

The surgeons also want to avoid ethical conflicts. They have occasionally used DBS to treat individual patients with other psychiatric and neurological syndromes, including Tourette's syndrome, epilepsy, minimally conscious state and cluster headaches. And they are thinking about obesity and drug addiction. But they know they must move cautiously.

The controlled OCD and depression trials should first be completed and analysed, says the neurosurgeons, to be absolutely sure that the therapy really works. They don't want to charge blindly on and risk failures and the attendant bad publicity. They also insist that progress in the field should be driven by psychiatrists rather than neurosurgeons, and closely controlled by ethics panels.

Ethicists worry about issues such as truly informed consent. They also debate the concept of personality, and whether it can be changed by DBS. "But patients don't see their obsessions and compulsions as part of their personality," comments Schormann. "They see them as something imposed on them, that they yearn to be rid of."

Patients who have been successfully operated on speak of regaining lives they had considered beyond hope. Monsieur F., one of Nuttin's OCD patients, was suicidal before DBS made his obsessive thinking and compulsive behaviour vanish within seconds. "Those who worry about ethics know nothing about it," he says. Herr Z., fresh off the operating table, hopes for the same freedom. He is a radio ham, and has recently learnt new languages to communicate with the world. He'd like to test out his new skills on a holiday, unconcerned by foreign germs. ■

Alison Abbott is Nature's senior European correspondent.

1. Benabid, A.-L. et al. *Acta Neurochir. Suppl.* **58**, 39–44 (1993).
2. Mayberg, H. et al. *Neuron* **45**, 651–660 (2005).
3. Nuttin, B. J. et al. *Neurosurgery* **52**, 1263–1272 (2003).



C. DARKIN

A new leaf

Record-keeping in the lab has stayed unchanged for hundreds of years, but today's experiments are putting huge pressure on the old ways. **Declan Butler** weighs up the pros and cons of electronic alternatives to that dog-eared notebook.

Pity the poor lab notebook. Almost everyone is publishing, sharing and searching information in electronic form, from biological and literature databases to blogs of their latest ideas. But lab notebooks are stuck in a time warp, still handwritten, on paper. Many scientists can barely understand their own scribbles from last week, let alone five years from now — as anyone who has had to decipher the hieroglyphics of a co-worker can testify.

We've been promised the 'paperless lab' since the early 1990s, but it hasn't taken off. Now, almost every laboratory is struggling with an explosion of data, much of it generated by automated instruments. An electronic version of the lab notebook can fully integrate with all these digital data and images. But most

researchers continue to make fuzzy printouts of gels and data tables, and glue them on to the leaves of their notebooks.

Things may finally be changing, however. E-notebooks are ready for prime time, says Douglas Perry, a laboratory informatics expert at the Indiana University School of Informatics in Indianapolis. New products are being rolled out across the pharmaceutical industry by a growing army of vendors, he says, noting that the industry's massive investment in e-notebooks has helped the technologies to mature.

So far, academia has been slower to respond. But labs that have switched from paper say e-notebooks have made life easier, improved quality, and revolutionized the way they share data and results.

At Sweden's Karolinska Institute, bench

scientists at the Centre of Excellence in Structural Genomics quickly adapted to their new paperless lab. They are using an e-notebook made by Stockholm-based Contur Technology. "It's intuitive, and easy to get started," says Johan Weigelt, the centre's chief scientist. "People got rid of their paper notebooks immediately."

Weigelt says the e-notebook lets scientists share everything: "the bioinformatics, expression data, sequence analysis and all the molecular biology, protein purification, crystallization and structural determinations".

Pål Stenmark, a Karolinska postdoc, says his co-workers find the e-notebook gives them easy access to each other's results. "You don't have to ask each other for protocols. You can read exactly what the other person did," he says.

Knowing that others regularly dip into his e-notebook has also prompted Stenmark to spruce up his own note-taking. "You have to be a bit more particular when writing; that's good because it's easy to be sloppy when writing a personal lab book." And when you can't find that crucial protocol? "It's also very nice to be able to search your own lab book."

Although the data entry 'pages' of an e-notebook look like a typical web form, they are dynamic not static. Once an experiment is set up, data is automatically captured and stored in a structured database, often alongside instrument settings. And where a compound, protein or gene is mentioned, it can be hyperlinked to structure and sequence data on the web.

You can add notes as you go along, compute or analyse results directly in your notebook, and store those in turn. Once the experiment is over, click on the 'submit experiment' button, and the computer gives it an indelible date and timestamp, and you sign it off with an encrypted digital signature. The pages can now never be modified by anyone, although comments can be added later, each with their own timestamp and signature.

It is this electronic security, and the need for detailed documentation to protect legal challenges to their inventions, that has driven drug companies to embrace e-notebooks. They also meet the US Food and Drug Administration's strict standards for record keeping for clinical trials and regulatory approvals, and may help discourage misconduct.

But the biggest advantage of going electronic should appeal to both industry and academia alike. With paper records, vast amounts of data lie unused and effectively hidden, whereas e-notebooks allow results to be instantly shared with collaborators. The pitfalls of paper are estimated to cost the drug industry alone over \$1 billion annually in lost opportunities and duplicated research.

However, installing new software on scientists' computers, and simply telling them to use it, is not enough. Most software needs to be customized to the working habits and research needs of individual groups, which means an investment of time and energy.

Experience shows that scientists won't adopt e-notebooks, if it means taking on additional computing tasks, unless it also makes their job markedly easier. Software has to adapt to the user, not the other way round. The Paris-based pharmaceutical giant Sanofi-Aventis spent years working with a French software provider, Klee Group, to develop a product (now available commercially as Kalabie) that met the specifications its scientists needed.

Cut to size

Customizing e-notebooks is proving more difficult in academia than in industry. Industry research groups are typically large and well-funded, with clearly defined project management and workflows, and generous IT support.

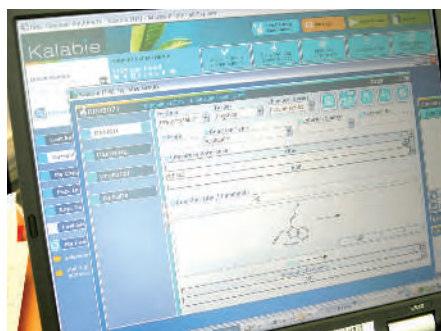
"Academics simply don't have funds for software that needs customizing to their situation, and there are almost as many different situations as there are labs," says Robert Cannon, a computational neuroscientist at Edinburgh University and founder of the Edinburgh-based notebook company Axiope.

And experimental protocols are continually changing. But redesigning the pages of e-notebooks, and adding fields and tables, usually entails modifying the structure of the underlying database. That requires the skills of database engineers, a luxury few labs enjoy.

Apart from a handful of big international collaborations that have database managers, such as CERN's Large Hadron Collider, most academic research is investigator-driven and

small-scale. Academic groups are often scattered across disciplines, and are perhaps overly creative in inventing different ways of working.

Academic e-notebooks need to include user-friendly tools allowing researchers to customize them without demanding computing skills, says Gwen Jacobs, who heads the open-source NeuroSys neuroscience e-notebook system at Montana State University, Bozeman. "Any system must be flexible enough to handle any type of data, and any experimental protocol," she says.



Note well: e-notebooks like SmartTea (top) and Kalabie are nudging paper records out of the picture.

"The nice thing about NeuroSys is that researchers can customize their templates and build user interfaces without a database administrator," says Maryann Martone, scientific coordinator of the Mouse Biomedical Informatics Research Network, based in La Jolla, California. Martone is carrying out an "exhaustive evaluation" of commercial and open-source e-notebooks.

Although most e-notebook vendors focus on the lucrative market in industry, Contur, Axiope and Rescendis in Columbus, Ohio, among others, are bringing out products targeting the needs of computer-illiterate academics.

And if you are keen to try one for free, you can download a basic open source Electronic Laboratory Notebook (<http://collaboratory.emsl.pnl.gov>). It lacks the cutting-edge database features of commercial products, but allows any group to share notebooks online and create templates for data entry. It also has a suite of data-analysis tools. The system was built by a group led by Jim Myers, a pioneer in e-notebooks at the Pacific Northwest National Laboratory in Richland, Washington.

Besides freeing academics from paper, elec-

tronic notebooks herald much deeper changes to the way science is practiced. Blogging has transformed communication on the web, and once scientists start making their e-notebooks available online — if only to remote collaborators — there will be a revolution in data sharing, predict observers.

Rather than spending time collecting their own data, scientists will organize themselves around shared data sets, much as astronomers do, says Monica Schraefel, a computer scientist at the University of Southampton, UK. Her SmartTea e-notebook project aims not just to get paper out of chemist's labs, but to get their experiments out to external collaborators as soon as they're finished.

Track records

Schraefel's group is also helping the scientists who barely use paper at all: bioinformaticians. "Whereas chemists have 400 years' practice using a lab book for their experiments, bioinformaticians, who do all their work on a computer, have none," she says.

In bioinformatics research, software is used to analyse data from big biology databases such as GenBank and Swiss-Prot. The 'results' are often strewn across thousands of computer files that the researchers struggle to organize.

"The strategies we've seen them invent to keep track of their work are amazing — and they fail," says Schraefel. "It gets to the point, they tell us, where it's easier to run an experiment again than to try to find the data. Great."

And if bioinformaticians can't keep track of the mess on their own hard drives, what hope do they have of sharing results with colleagues? Schraefel is tackling this problem by building unobtrusive software dubbed MyTea, which runs in the background and tracks related files and notes. The idea, she says, is to create "representations of work in progress that can be shared" — in short, lab notebooks.

Whatever form e-notebooks take, some are concerned about how robust archived copies will be. "I'm very nervous about the fragility of digital information," says Martone. "We can still read papyrus and stone monuments from over 2,000 years ago. We sometimes can't read digital files from 10 years ago."

But Stenmark points to the "disastrous" risk of losing a paper notebook, perhaps in the train on the way home from a lab party. He sleeps easier knowing his e-notebook is automatically backed up nightly to multiple locations.

And those attached to their dog-eared lab books shouldn't worry: even the drug industry is reluctant to abandon paper completely. Although electronic digital signatures and time stamps are now legally accepted, they haven't yet been tested in a court case. So companies that have gone electronic are playing it safe. After each experiment is submitted to the e-notebook, managers print it out, have the researcher and a witness sign each page using a pen, and lock the lot in a safe.

Declan Butler is a senior reporter at Nature.

M. C. SCHRAEFEL

KLEE GROUP

BUSINESS

Silicon down to the wire

Microchip-makers are starting to look beyond silicon, and what they see, reports **Colin Macilwain**, is a semiconductor industry of a very different complexion — but not for some time yet.

Jack Kilby, the Texas Instruments engineer who invented the integrated circuit in 1958 and won the Nobel Prize in Physics for it in 2000, died of cancer last month. But the silicon-chip technology his idea gave birth to will be ubiquitous for a long time yet. That's the emphatic message of the latest roadmap for the semiconductor industry, which will be released in San Francisco on 13 July.

But the authors of the 2005 International Semiconductor Technology Roadmap, which will be previewed at the Semicon West 2005 trade show, have added extensive treatments of 'emerging technologies' that could work with, or even rival, the silicon-based CMOS (complementary metal-oxide semiconductor) technology that's got the industry where it is today.

The roadmap — a venture involving about 1,000 semiconductor specialists worldwide — is revised every two years, and the latest version reflects the chip-makers' expectation that CMOS will reach absolute limits on its performance by around 2020.

"The roadmap is giving a good deal of attention to alternatives," explains Ralph Cavin, an engineer with Semiconductor Research Corporation (SRC) at Research Triangle Park in North Carolina and a member of the roadmap steering committee. "But you should not be misled — CMOS is going to be around for a long time." For the logic circuits that form the heart of computers and other electronic devices: "we haven't found anything that's superior to CMOS," he says. But in memory devices, he believes, there may be viable alternatives.

Memory chips sit at the worthy but dull end of electronics, however. And the roadmap doesn't dodge the question of which new technologies could find their way into the computer's brain — the logic circuit. But its exhaustive examination of the issues raises as many questions as it answers.

The "emerging research architecture" section of the document, which deals with technologies that could be used in logic circuits, looks in detail at several approaches, most notably quantum computing, biology-based approaches and cellular arrays.

The quantum computer is probably the alternative that the public hears most about. But Victor Zhirnov, a physicist-turned-electrical

engineer at SRC and member of the roadmap's working group on emerging devices, takes a dim view of its realistic prospects.

Over the past five years, there have been vastly more research papers published on quantum computing than on the other alternatives. "Ninety per cent of these papers are theoretical," says Zhirnov. "And their assumptions are questionable — for example, that the device will be 'isolated from its environment.'" They might as well, Zhirnov caustically observes, assume movement faster than light.

The last roadmap, in 2003, was equally scathing about the prospects for quantum computers. They depend on wave functions that "would easily decohere when interacting with an external environment", it said. "Although enormously capable for a few

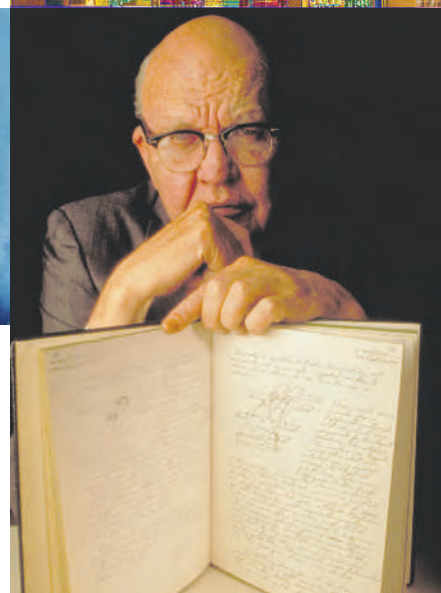
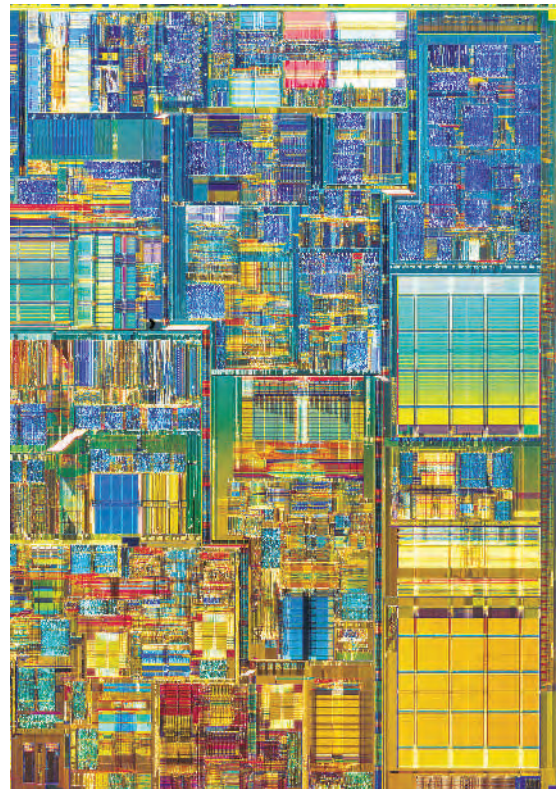


Then: the first integrated circuit (above), invented by Jack Kilby (right), and now: Intel's Pentium 4 microprocessor (top right).

selected algorithms, such as encryption or deep database searching, quantum computing is not seen yet as being of more general interest." This time, the approach will get less play, says Jim Hutchby, another physicist-turned-engineer at SRC, who chairs the group that is writing the roadmap.

The other options all have their own flaws, however. Zhirnov puts most faith in approaches inspired by biology, marvelling at the phenomenal processing power — and thrifty fuel consumption — of the human brain. "I personally believe we need to focus on how the brain works," he says, "but our understanding of its architecture is very embryonic."

He points to work suggesting that each neuron works not as a binary switch, but as a server



TEXAS INSTRUMENTS; INTEL

in its own right (see C. Koch, *Nature* **385**, 207; 1997). Zhirnov hopes the roadmap will help emphasize the need for an interdisciplinary effort. "There's a lack of communication between engineers and neuroscientists", he says.

Another approach for some computer functions would be cellular arrays, which involve interactions between a large grid of nano-scale elements — electrostatically coupled quantum dots, perhaps, or coupled magnets. The roadmap authors see potential for this

approach in some specialized applications such as character recognition.

The roadmap will also explore the vexed question of how alternatives to the silicon chip will tolerate defects. The stunningly high reliability of solid-state silicon circuitry is "probably the most important aspect of CMOS", says Texas Instruments' Bob Doering, co-chair of the roadmap steering group. But the reliability of alternatives is far less sure. According to the 2003 document, each working nanoscale device might need between 1,000 and 10,000 back-ups to make the whole circuit work reliably — although that point is disputed and may be toned down this year.

Then there's heat dissipation. Getting the heat out of today's CMOS microprocessors is already a huge challenge. And Cavin says that CMOS will soon operate close to the theoretical limit that could be achieved by any alternative that relies on moving electrical charge.

The roadmap is a unique global exercise, in which bitter competitors lay aside their differences to reach agreement on where the technology of their business is heading. It originated in 1994 with Sematech in Austin, Texas, a collaboration of US semiconductor manufacturers with their backs to the wall at the time. Now it is an international effort involving more than 1,000 engineers and physicists.

But the authors of the roadmap don't see it as their job to pick winners between rival approaches. Instead, Doering explains, "its main function is to raise red flags" that engineers can get working on.

The industry does, however, need to identify which technologies to support. With the 2020 deadline in mind, it is working with the US National Science Foundation (NSF) on a nanoelectronics initiative to support students and projects on the emerging approaches. Given the time it will take to get these up and running, semiconductor-makers want to know which ones to back by 2008, says Michael Roco, the NSF official behind the initiative.

But this research effort doesn't anticipate finding new technologies that will allow chips to extend the amazing run of ever-greater performance and miniaturization that CMOS has enjoyed — performance has doubled every 18 months for decades. "The focus will no longer be on miniaturization or speed," says Roco. "It will be on new functions. It will be devices that you put on your skin to monitor your health, or ways to connect between a neurological system and a machine."

To open up this brave new world, "there need to be some decisions made", says Hutchby. "What matters now is not density or clock speed," he adds. "It is new functionality that CMOS was not going to address." ■

IN BRIEF

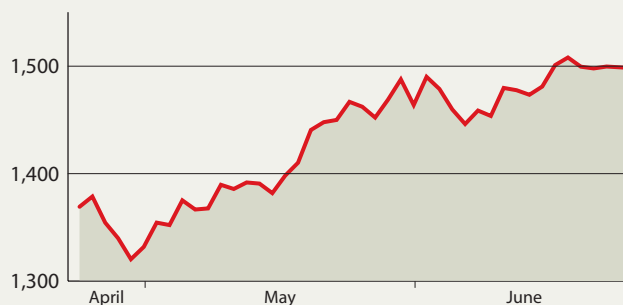
STEM-CELL FLOAT Edinburgh-based research company Stem Cell Sciences says it will seek a listing on London's Alternative Investment Market later this month. The company wants to raise £10 million (US\$18 million) by the proposed stock-market float. Stem Cell Sciences, which was founded in 1994 in Australia by biologist Peter Mountford, employs 30 people and has research interests in Japan as well as in Scotland and Australia. If the public offering succeeds, the company will join a small handful of other listed European companies with an interest in stem-cell research.

CHANGE OF PLAN Genentech, the California biotechnology company, is paying \$408 million to take over a pharmaceutical factory left idle as the result of a drug suspension. Biogen Idec of Cambridge, Massachusetts, had planned to produce the multiple sclerosis drug Tysabri at its facility in Oceanside, California, but had to withdraw the drug in February because of safety concerns. The company says it is taking a \$50-million loss on the sale. Genentech, which has research collaborations with Biogen Idec, says it will start making Avastin — a treatment for colon cancer — at the 430-employee plant in 2007.

BIG CARBON EXCHANGE Two of Europe's main markets for carbon emissions have announced plans to merge. The combination of Paris-based Powernext Carbon and the European Climate Exchange (ECX), which is based in Amsterdam and does most of its business in London, is expected to create the continent's largest emissions exchange. Analysts expect more consolidation between the half-a-dozen existing exchanges in Europe as trading in carbon dioxide emissions heats up.

MARKET WATCH

Nanotechnology stocks



Stocks in companies related to nanotechnology have rebounded this spring, as investors grow a little less risk-averse.

The Lux Research Nanotechnology index, which covers a global cross-section of companies that supply nanotechnology products, build nanotech tools or rely heavily on the use of the technology, has grown by more than 9% over the past two months. And the biggest factor at work, says Peter Hebert, president of the New York-based consultancy, is investors' renewed willingness to put their money into small, technology-based companies.

But strong individual performances by some of the 25 companies in the index also played a role. Strongest of all was French company Flamel Technologies, of Lyon, a specialist in nanoparticle-based drug-delivery systems, whose stock rose from \$13 to \$20 on news of a management shake-up.

Another big winner was Canadian company Westaim, whose Massachusetts-based subsidiary Nucryst makes silver nanoparticles for wound dressings. Nucryst stock rose from \$2.40 to more than \$3 on strong sales reports.

And stock in Symyx Technologies, a specialist in nanoscale catalysts based in Santa Clara, California, went from \$22 to \$27 on news of deals with Exxon and Dow, who will use its products in petroleum processing and plastics production, respectively.

Hebert predicts that nanotech stocks — whose performance pipped that of the Nasdaq technology index for the two-month period — could do better yet. "The summer is looking very strong," he says, noting that nanotech companies could be exempt from adverse factors, such as weak demand for semiconductors, that could hold back the technology sector as a whole.

► www.luxresearchinc.com

SOURCE: LUX RESEARCH

Call for openness about farm-animal experiments

SIR — The recent call by the Nuffield Council on Bioethics (“UK panel urges animal researchers to go public” *Nature* 435, 392; 2005) for scientists to discuss more openly the use of animals in experiments should start with farm animals. During their lives, most commercial livestock are subject to experimentation at several levels: the individual farm; the national flock or herd; and the global livestock industry. Indeed, at a more basic level, a farm’s day-to-day and year-to-year refinement of animal-management practices and business activities is itself scientific, experimental and involves animals.

“Commercial livestock are subject to experimentation at several levels: the farm, the national flock or herd and the global livestock industry.”

— Ian G. Colditz

At the individual farm level, genetic improvement programmes provide one example of animal experimentation through collection of data on pedigree and on individual performance such as body weight and milk production. Another example is collection of blood for serology, on the farm or at the point of slaughter, to test hypotheses on disease epidemiology.

What benefits might flow from a broader appreciation of commercial farms’ dependence on farm-based animal experimentation within their enterprise and more broadly across their industry?

First, experimental practices, especially genetic improvement programmes, can have welfare consequences that deserve our attention (see W. M. Rauw *et al. Livest. Prod. Sci.* 56, 15–33; 1998). These can be beneficial, for example when animals are selected for calving ease. They can also be harmful, for example in broiler chickens whose legs are weakened by selection for fast growth rate.

Second, the discussion may help illuminate the nature of human–animal relationships and potentially reduce the stigma associated with use of animals in universities and research institutes.

Ian G. Colditz

CSIRO Livestock Industries, Locked Bag 1, PO, Armidale, New South Wales 2350, Australia

Plagiarism criteria ignore the way research evolves

SIR — Your Special Report on plagiarism, “Taking on the cheats” (*Nature* 435, 258–259; 2005), does not, in my opinion, appreciate

the way in which scientific research evolves.

In my experience, incremental progress is often reported at one or more scientific conferences, until a comprehensive manuscript can be submitted to a reputable journal for publication. The work may be further published in research monographs, in review articles and, on occasion, in textbooks.

It seems to me a misunderstanding to insist that a piece of work must be published only once.

A series of progress reports would naturally have extended sections in common — if this were not allowed, no scientist would consider presenting work at a conference with published proceedings. Increasingly, conference papers are being published as regular books and sometimes even as special issues of standard journals. Is it not defeating the scientific purpose of conferences if final papers are expected, rather than discussion papers on work in progress?

Similar conflicts can arise between journal and book publishers. I submitted a paper to a scientific journal and later incorporated a description of it in a research monograph. However, the efficient book publisher got the monograph on the street two months after receiving the manuscript, whereas the journal turned out to have a backlog resulting in accepted papers waiting more than a year to appear in print. In this instance, could I be accused of committing ‘self-plagiarism’ on the basis of overlaps, when I have no control over the schedules of the publishers?

Bent Sørensen

Department of Mathematics and Physics, Roskilde University, Universitetsvej 1, bld. 27, DK-4000 Roskilde, Denmark

Six-word rule could turn description into plagiarism

SIR — Your Special Report on the software that journal editors are considering to help them catch academic cheats (*Nature* 435, 258–259; 2005) suggests that six words used contiguously in more than one published paper now constitutes plagiarism. I design novel oligonucleotides to inhibit STAT3 activity in hormone refractory prostate cancer. I use this phrase of more than six words to describe what I do in the introduction to my manuscripts. Am I committing self-plagiarism?

Plagiarism must absolutely be defined not by words used but by data shown. That is the serious offence, not that someone reuses a key descriptive phrase in several papers.

Beverly E. Barton

Department of Surgery/Division of Urology, UMDNJ-NJMS MSB G519, 185 South Orange Avenue, Newark, New Jersey 07103, USA

Penalties plus high-quality review to fight plagiarism

SIR — Your Special Report (*Nature* 435, 258–259; 2005) on plagiarism in scientific texts overlooks an important related problem, namely multiple publications of the same data (graphs or pictures).

The computer programs for identifying duplicate text in manuscripts do not seem capable of handling this particular kind of dishonesty. In one case I have identified, the authors had published the same electron micrographs repeatedly in different journals, and attempted to do so with a manuscript that was sent to me for review. In trying to hide self-plagiarism, the copied figures were presented upside down, rotated by 90°, or slightly cut or expanded. This is reminiscent of Jan Hendrik Schön’s scientific misconduct (*Nature* 417, 367–368; 2002).

When I informed the editor, she rejected the paper with a letter stating that it was unacceptable, as a reviewer had detected the plagiarism. But there are good arguments for journals to go further and take punitive measures, such as notifying the authors’ institution and, if applicable, their funding agency. Papers identified as fraudulent after publication should be removed from the web version of the journal and be replaced by a note disclosing the scientific misconduct. The authors’ consent should not be required.

Even more serious sanctions may need to be discussed publicly and included in the guidelines for manuscript submission. For example, the authors could be banned from publishing, permanently or temporarily, at least in the cheated journal. Funding agencies could add pressure by stating that scientists practising scientific misconduct will be denied the right to submit proposals.

Last, but not least, many cases of scientific misconduct could be avoided if the work of good reviewers were appreciated more explicitly. I tell myself that I’m doing it for the benefit of science. But, lacking appreciation, some reviewers might do their job in a sloppy manner. Why not encourage them? The only journal I know that makes an award for excellence of review is *Environmental Science and Technology*. Publishing a list of the top 10 or 20 reviewers each year may encourage more in-depth evaluations of manuscripts.

Klaus Wittmaack

Institute of Radiation Protection, GSF National Research Centre for Environment and Health, 85758 Neuherberg, Germany

Nature’s policy on duplicate publications is outlined at www.nature.com/nature/authors/policy/index.html#a4. Nature has on occasion notified employers concerning misconduct detected during peer review, and reserves the right to bring it to readers’ attention — Editor, *Nature*.

BOOKS & ARTS

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

The race for the bomb

How close was Nazi Germany to developing atomic weapons?

Hitlers Bombe: Die Geheime Geschichte der Deutschen Kernwaffenversuche

by Rainer Karlsch

Deutsche Verlags-Anstalt: 2005. 432 pp.
€24.90, SFr 43.50

Dieter Hoffmann

Berlin historian Rainer Karlsch's book deals with one of the most controversial questions in the modern history of science in Germany: did the nation make serious progress towards developing an atomic bomb? The sensational title of his book, *Hitlers Bombe*, could suggest that German scientists built and tested an atomic bomb, but this implication is not borne out by the book's content. Germany certainly did not have an atomic bomb the size of those that the United States dropped on Hiroshima and Nagasaki in August 1945, which is today's standard meaning of the term 'atomic bomb'.

So what does Karlsch mean when he uses the word 'bomb'? He uses no unique nomenclature, speaking of tactical atomic weapons, of nuclear bombs, of thermonuclear hollow

explosive bombs, and of uranium-235 bombs. In fact he tells us a story of how Germany developed a small nuclear weapon, which he says was tested at least twice.

The physical principle of the bomb is never described precisely or comprehensively. For example, Karlsch describes a design consisting of two spheres, nested one inside the other. The smaller one in the centre contained heavy water, surrounded by a sphere of nuclear fuel. The outer spherical surface was covered in conventional hollow explosive charges, which, when ignited, compressed the heavy water so intensely that it started a fusion reaction. The neutrons generated in this way would have triggered fission reactions in the nuclear fuel. It is plausible that the Germans based their work on advanced research into hollow explosive charges, but even so, by ordinary physics, the reaction speeds, pressure and available temperature are at least two orders of magnitude too small to initiate a fusion reaction. Furthermore, it is unclear how the Germans obtained their plutonium or their enriched uranium. Karlsch's explanation, both here and

elsewhere, is vague, and he generally gives only qualitative descriptions without concrete figures. So it is not clear how well this thermonuclear hollow explosive bomb, as well as other bomb designs, would have worked, if it worked at all.

Nevertheless, Karlsch believes that a nuclear weapon could first have been tested in October 1944 in northern Germany, on the island of Rügen. A second test reportedly killed several hundred prisoners at a concentration camp in Thuringia in March 1945. But these tests bring us to another problem of the book: the historical reliability of Karlsch's sources. There is doubt over the evidence for the first test in Rügen, as the first report about the event, shortly after the war, is controversial and unreliable. The second event is better supported by documentary evidence. It seems certain that some kind of new and powerful weapon was tested in Thuringia in March 1945 — but what sort of weapon is not clear.

Karlsch attempts to prove his hypothesis that it was a nuclear weapon with data from

soil probes, analysed with methods from modern nuclear physics. The first results of such analyses are still inconclusive, although more measurements, by independent institutes, are under way. So Karlsch can only back his hypothesis with eyewitness reports. Such testimony is notoriously problematic (especially when it is given much later) and is often highly contradictory. As a result, the central argument of Karlsch's book is not all that convincing and is in key parts inconclusive. There is no precise and physically plausible description of the bomb's design, and no reliable analysis of the purported test region to demonstrate that there really was a nuclear reaction.

Even so, Karlsch has written for the most part an interesting, even valuable, book and demonstrates his credentials as a serious historian. The book makes clear that it was well known in the German scientific community that uranium and other nuclear fuels, and even nuclear fusion, could be used to make powerful new weapons. Evidence for this comes from archival material that Karlsch has collected, much of it previously unknown. For example, he discovered a wealth of material in the Russian archives, confiscated from Germany by the Red Army in 1945, including a patent by Carl Friedrich von Weizsäcker (dated summer 1941) for energy production using a uranium pile; this also suggests that plutonium was used as a nuclear fuel.

These documents, and other evidence amassed by Karlsch, demolish the widespread myth that Germany had no chance of building an atomic bomb during the war, and that, even if it did, German physicists (notably Werner Heisenberg's group) would have done whatever was necessary to make sure that such a terrible weapon never made it into the hands of the Nazis. On the contrary, Karlsch demonstrates that numerous German scientists and engineers were grappling with the problem of developing nuclear weapons.

Previous research has focused on Heisenberg and his team, but Karlsch has shifted the focus to other groups. For instance, he shows that Kurt Diebner was perhaps the central figure in Germany's attempt to make an atomic bomb. Not only did Diebner have an idea of the proper arrangement of the uranium cubes in the pile, but Karlsch has uncovered circumstantial evidence that the Diebner group had a critical pile, although not for long enough to produce a significant amount of plutonium. Karlsch also shows that other groups were developing ideas for a nuclear weapon.

Another interesting story related for the first time by Karlsch concerns the activities of Walther Gerlach. He became the administrative head of German nuclear research in 1943, but has never previously been thought to have played a central role in the development of an atomic bomb; he is usually seen only as an ally of Heisenberg and his group. In contrast, Karlsch shows (with the help of new archival material and diligent checking of the old) that

Gerlach was one of the driving forces in the development of a nuclear weapon and was a powerful patron of Diebner and his efforts. This could explain why in 1945, while held at Farm Hall near Cambridge, UK, with other leading German nuclear physicists, Gerlach had the air of a defeated general and had a nervous breakdown after the Hiroshima bomb.

Karlsch's central argument — that Germany

had developed and tested a nuclear bomb — remains more a sensationalist construct than a proven fact. But his book presents a wealth of interesting and valuable information about the attempt in Nazi Germany to develop a nuclear weapon. ■

Dieter Hoffmann is at the Max Planck Institute for the History of Science, Wilhelmstrasse 44, Berlin 10117, Germany.

A happy gathering

Happiness: Lessons From a New Science

by Richard Layard

Allen Lane/Penguin: 2005. 320 pp.
£17.99/\$25.95

Making Happy People: The Nature of Happiness and its Origins in Childhood

by Paul Martin

Fourth Estate: 2005. 306 pp. £15.99

Happiness: The Science Behind Your Smile

by Daniel Nettle

Oxford University Press: 2005. 224 pp.
£9.99, \$21

Dylan Evans

Books about the scientific study of happiness, it seems, are rather like buses: you spend ages waiting for one, and then three come along at once. The same trend is evident in the science itself, as researchers rush to remedy a long-standing deficit. For much of the twentieth century, psychologists paid scant attention to happiness and related notions, but in the past decade it has suddenly become a hot topic. In this respect, psychology seems to be returning to its roots, as happiness was a central concern for many of the field's founding fathers, such as William James and Sigmund Freud.

The three latest books on the subject have several things in common. They are all popular summaries of the field, aimed at the general public, rather than scientific monographs

aimed at specialists. They are all well written, accurate and engaging. And they all cover broadly similar ground. For example, they all start by discussing the various different meanings of happiness and the ways in which happiness can be measured. They all go on to discuss the main factors that make people more or less happy, including money, life events, personality and genes. All explore the increasing evidence for the idea that being happy is good for your health. And they all make the point that scientific research often contradicts our commonsense intuitions about how best to obtain happiness.

That said, there are also differences in the general approach. In *Happiness: Lessons From a New Science*, Richard Layard examines how research can inform social policy, and argues that happiness is a more sensible goal for society than economic growth. In *Making Happy People*, Paul Martin is more concerned with the implications of the research for parenting and education. Daniel Nettle, in his book *Happiness: The Science Behind Your Smile*, prefers to stick to the science itself, and is less concerned with its applications to practical contexts; this makes him — rightly, I think — more sceptical of the idea that happiness is the ultimate goal of human life. In their rush to apply the scientific research to practical matters, Layard and Martin both champion a rather crude version



The smiling faces of children in Baghdad suggest that wealth is not directly related to happiness.

SHEHZAD NOORANI/STILL PICTURES

EXHIBITION

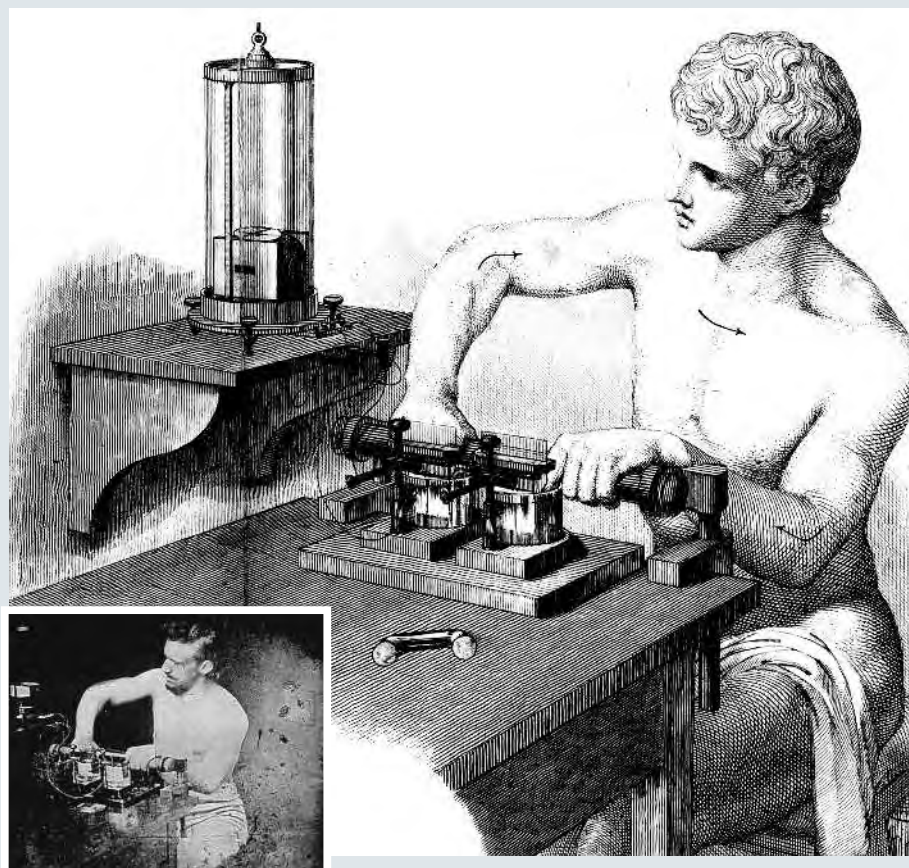
Apollo in the lab

Why is Apollo at the bench? This textbook engraving was designed by the nineteenth-century natural scientist Emil Du Bois-Reymond, a founder of electrophysiology. It was based on a photograph he took of his brother (inset).

Du Bois-Reymond was well schooled in the classics and was a member of a new school of scientists, which held that biological tissues were subject to the laws of physics. He also believed that experimentation has its own aesthetics.

He worked closely with instrument-makers, for example in developing his *Multiplikator*, a precision galvanometer that could detect the small transient currents in human muscles when flexed. He believed that the researcher must, through athletic training, become an experimenting Apollo, a part of his instrument. He himself worked out in a makeshift gym in his home where he originally had his laboratory. He became the director of the Physiological Institute in Berlin, but he was unhappy at having less day-to-day control of the experimental work carried out there.

An exhibition of the life and work of Du Bois-Reymond can be seen at the Berlin Museum of Medical History at the Charité Hospital until 2 October. A.A.



BERLIN MUSEUM OF MEDICAL HISTORY

of utilitarianism, although neither provides much in the way of an argument for this. Nettle's position is more sophisticated, as he allows room for a range of other human goods alongside happiness, such as "purpose, community, solidarity, truth, justice, and beauty", which cannot simply be converted into some imaginary universal currency called utility.

Layard is the best guide to the complex relationship between happiness and money, although this is also well analysed by Martin and Nettle. Drawing on recent work by economists such as Robert Frank, Layard presents an array of graphs and tables showing that rising affluence in the developed world has not increased average levels of happiness. Indeed, there is some evidence that people in the developed world have actually become less happy as they have got richer, at least in some respects. All three books explore the reasons for this apparent paradox, but only Nettle provides something approaching a deep explanation. He proposes that natural selection has endowed us with an implicit theory about what makes us happy that is false by design. In other words, unhappiness is not always a sign that our psychological mechanisms have gone wrong. On the contrary, "the wanting system is supposed to enslave you, to make you maximise your reproductive success". Our tendency to be mistaken in our beliefs about what will make us happy is, Nettle explains, "a particularly cruel trick played by

our evolved mind to keep us competing".

Martin is at his best when discussing how the education system so often fails to equip children to lead happy lives, and how it might be changed to remedy this deficit. He makes a powerful case for happiness to feature prominently on the educational agenda, and this is a welcome antidote to the narrow view of education as a preparation for the workplace that is becoming prevalent in many Western countries. His book should be required reading for anyone working in education policy.

None of these three authors can resist the temptation to offer practical tips on how to be happy. But it is a great relief that they all avoid the more messianic tones that have blighted some of the offerings of the 'positive psychology' movement launched by the psychologist-turned-guru Martin Seligman.

If I had to recommend just one of these books, it would be Nettle's, because it conveys about the same amount of information as the other two books in about half the number of words. Yet the conciseness is achieved with a lightness of touch that makes it a delight to read. And Nettle is more aware than Layard and Martin of the paradoxes inherent in the pursuit of happiness — paradoxes that so often make happiness such an elusive goal. ■

Dylan Evans is a senior lecturer in intelligent autonomous systems at the University of the West of England, Frenchay Campus, Coldharbour Lane, Bristol BS16 1QY, UK.

NEW IN PAPERBACK

Mutants: On Genetic Variety and the Human Body

by Armand Marie Leroi (Penguin, \$16)

"*Mutants* is an exquisitely life-enhancing book. It captures what we know of the development of what makes us human, and it recognizes the random tragedy inflicted by nature and nurture." Peter Little *Nature* **427**, 101-102 (2004).

A Brief History of the Human Race

by Michael Cook (Granta, £9.99)

"An elegant, quick and engaging way to review what has happened in history, to learn much that is new, and to appreciate the past of the whole world, not just the West. It meets scientists almost halfway, trying to ground the events of history literally in the material facts of the planet." Melvin Konner *Nature* **428**, 123-124 (2004).

Sight Unseen

by Melvyn Goodale and David Milner (Oxford University Press, £14.99)

"Goodale and Milner emphasize that much of what goes on in our brains, and even in our cortices, escapes our conscious 'I', partly because of the separation of the visual systems for perception and action... this volume is a perfect present for anyone even remotely interested in the brain." Manfred Fahle *Nature* **429**, 703 (2004).

The mental Universe

The only reality is mind and observations, but observations are not of things. To see the Universe as it really is, we must abandon our tendency to conceptualize observations as things.

Richard Conn Henry

Historically, we have looked to our religious leaders to understand the meaning of our lives; the nature of our world. With Galileo Galilei, this changed. In establishing that the Earth goes around the Sun, Galileo not only succeeded in believing the unbelievable himself, but also convinced almost everyone else to do the same. This was a stunning accomplishment in 'physics outreach' and, with the subsequent work of Isaac Newton, physics joined religion in seeking to explain our place in the Universe.

The more recent physics revolution of the past 80 years has yet to transform general public understanding in a similar way. And yet a correct understanding of physics was accessible even to Pythagoras. According to Pythagoras, "number is all things", and numbers are mental, not mechanical. Likewise, Newton called light "particles", knowing the concept to be an 'effective theory' — useful, not true. As noted by Newton's biographer Richard Westfall: "The ultimate cause of atheism, Newton asserted, is 'this notion of bodies having, as it were, a complete, absolute and independent reality in themselves.'" Newton knew of Newton's rings and was untroubled by what is shallowly called 'wave/particle duality'.

The 1925 discovery of quantum mechanics solved the problem of the Universe's nature. Bright physicists were again led to believe the unbelievable — this time, that the Universe is mental. According to Sir James Jeans: "the stream of knowledge is heading towards a non-mechanical reality; the Universe begins to look more like a great thought than like a great machine. Mind no longer appears to be an accidental intruder into the realm of matter... we ought rather hail it as the creator and governor of the realm of matter." But physicists have not yet followed Galileo's example, and convinced everyone of the wonders of quantum mechanics. As Sir Arthur Eddington explained: "It is difficult for the matter-of-fact physicist to accept the view that the substratum of everything is of mental character."

In his play *Copenhagen*, which brings quantum mechanics to a wider audience,

Michael Frayn gives these words to Niels Bohr: "we discover that... the Universe exists... only through the understanding lodged inside the human head." Bohr's wife replies, "this man you've put at the centre of the Universe — is it you, or is it Heisenberg?" This is what sticks in the craw of Eddington's "matter-of-fact" physicists.

Discussing the play, John H. Marburger III, President George W. Bush's science adviser, observes that "in the Copenhagen interpretation of microscopic nature, there are neither waves nor particles", but then frames his remarks in terms of a non-existent "underlying stuff". He points out that it is not true that matter "sometimes

Physicists shy from the truth because the truth is so alien to everyday physics. A common way to evade the mental Universe is to invoke 'decoherence' — the notion that 'the physical environment' is sufficient to create reality, independent of the human mind. Yet the idea that any irreversible act of amplification is necessary to collapse the wave function is known to be wrong: in 'Renninger-type' experiments, the wave function is collapsed simply by your human mind seeing nothing. The Universe is entirely mental.

In the tenth century, Ibn al-Haytham initiated the view that light proceeds from a source, enters the eye, and is perceived. This picture is incorrect but is still what most people think occurs, including, unless pressed, most physicists. To come to terms with the Universe, we must abandon such views. The world is quantum mechanical: we must learn to perceive it as such.

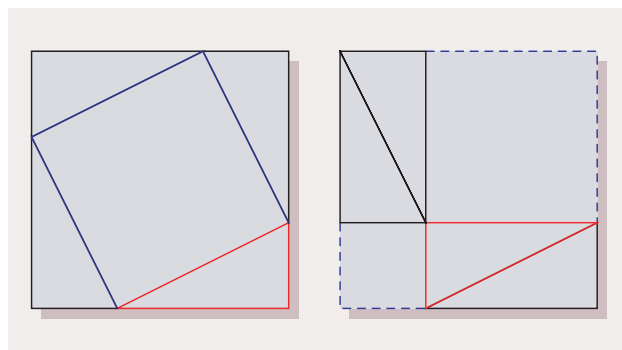
One benefit of switching humanity to a correct perception of the world is the resulting joy of discovering the mental nature of the Universe. We have no idea what this mental nature implies, but — the great thing is — it is true. Beyond the acquisition of this perception, physics

can no longer help. You may descend into solipsism, expand to deism, or something else if you can justify it — just don't ask physics for help.

There is another benefit of seeing the world as quantum mechanical: someone who has learned to accept that nothing exists but observations is far ahead of peers who stumble through physics hoping to find out 'what things are'. If we can 'pull a Galileo', and get people believing the truth, they will find physics a breeze.

The Universe is immaterial — mental and spiritual. Live, and enjoy. ■

Richard Conn Henry is a Professor in the Henry A. Rowland Department of Physics and Astronomy, The Johns Hopkins University, Baltimore, Maryland 21218, USA.



Proof without words: Pythagoras explained things using numbers.

behaves like a wave and sometimes like a particle... The wave is not in the underlying stuff; it is in the spatial pattern of detector clicks... We cannot help but think of the clicks as caused by little localized pieces of stuff that we might as well call particles. This is where the particle language comes from. It does not come from the underlying stuff, but from our psychological predisposition to associate localized phenomena with particles."

In place of "underlying stuff" there have been serious attempts to preserve a material world — but they produce no new physics, and serve only to preserve an illusion. Scientists have sadly left it to non-physicist Frayn to note the Emperor's lack of clothes: "it seems to me that the view which [Murray] Gell-Mann favours, and which involves what he calls alternative 'histories' or 'narratives', is precisely as anthropocentric as Bohr's, since histories and narratives are not freestanding elements of the Universe, but human constructs, as subjective and as restricted in their viewpoint as the act of observation."

FURTHER READING

Marburger, J. *On the Copenhagen Interpretation of Quantum Mechanics*
www.ostp.gov/html/Copenhagentak.pdf (2002).
 Henry, R. C. *Am. J. Phys.* **58**, 1087–1100 (1990).
 Steiner, M. *The Applicability of Mathematics as a Philosophical Problem* (Harvard Univ. Press, Cambridge, MA, 1998).

NEWS & VIEWS



NEUROSCIENCE

A home for the nicotine habit

Julie A. Kauer

Nicotine is extremely addictive, but it can also improve cognitive performance. Attempts to unravel the complex pathways underlying these effects pinpoint a single type of receptor in just one brain region.

Neuroscience is a somewhat unusual field, as it studies the workings of a single organ from microscopic to holistic levels. But a major technical challenge is how to link the two ends of the spectrum — how do specific molecules in the brain control behaviour? In this issue, Changeux and colleagues (Maskos *et al.*, page 103)¹ find that just one subunit of a neurotransmitter receptor, when active in a tiny region of the brain, is responsible for characteristic behavioural responses to nicotine in mice.

The addictive effects of nicotine are notorious. Smoking-related disease is responsible for many preventable deaths each year — up to 20% of mortality in developed countries alone². Nicotine binds exclusively to receptors called nicotinic acetylcholine receptors (nAChRs) on the surface of neurons in the brain. Normally, these receptors are activated by the neurotransmitter acetylcholine, and loss of acetylcholine-releasing neurons is implicated in Alzheimer's disease. Moreover, nicotine enhances cognitive performance. So understanding the molecular interactions of nicotine with its receptors may have significant benefits for human health, and provide clues to normal cognition.

Nicotinic acetylcholine receptors consist of five subunits, of which there are 16 varieties. The numerous resulting receptor combinations expressed throughout the nervous system thus potentially mediate the various

behavioural effects of nicotine and its natural counterpart acetylcholine.

In the brain, nAChRs containing the $\beta 2$ subunit are most prevalent, so in 1995 Changeux's laboratory produced genetically engineered mice lacking this subunit (termed $\beta 2^{-/-}$ mice) to discover how it contributes to normal function^{3,4}. The mice showed mild learning impairment in certain tasks (but not in others), supporting a role for $\beta 2$ nAChRs in cognition⁴. Furthermore, whereas normal mice rapidly learned to self-administer nicotine by pressing a lever, the $\beta 2^{-/-}$ mice did not, implicating these receptor subunits in mediating the reinforcing or rewarding properties of nicotine⁵. But such studies are imperfect: a behavioural deficit may result not from the absence of a certain molecule in the adult, but from abnormal development at an earlier stage. Moreover, the presence of $\beta 2$ -containing receptors throughout the brain made it difficult to assign altered behaviours in the $\beta 2^{-/-}$ animals to specific brain pathways.

Changeux's laboratory has now addressed¹ these issues by judicious reintroduction of the $\beta 2$ subunit into a specific brain region of $\beta 2^{-/-}$ mice. The midbrain ventral tegmental area (VTA) is strongly implicated in the response to natural rewards, such as food or sex, as well as the reinforcing effects of various drugs of abuse. All addictive drugs, for example, elicit release of the neurotransmitter dopamine, which is manufactured in neurons

of the VTA, and rodents will self-administer nicotine (or cocaine or morphine) directly into this region. Changeux and colleagues injected a virus encoding the $\beta 2$ nAChR subunit directly into the VTA of $\beta 2^{-/-}$ mice. In the following few days, the $\beta 2$ subunit began to be expressed in VTA neurons, forming functional receptors with other nAChR subunits already expressed there.

In normal mice, nicotine excites dopamine-containing cells in the VTA, resulting in dopamine release at nerve terminals; in $\beta 2^{-/-}$ mice, these responses are lost. But when the $\beta 2$ subunit was reintroduced into the VTA of $\beta 2^{-/-}$ animals, both responses to nicotine were restored. How does this translate into behaviour? Whereas $\beta 2^{-/-}$ animals did not self-administer nicotine, remarkably, $\beta 2^{-/-}$ mice with the reintroduced $\beta 2$ gene did — and did so nearly as often as control animals, suggesting that the rewarding effects of nicotine are entirely restored by local reintroduction of this receptor subunit. Given the intricacies of the brain, it is striking that reintroduction of a single molecule to just one small area of the brain should so dramatically affect behaviour.

One issue still to be addressed is whether $\beta 2$ -treated $\beta 2^{-/-}$ animals will self-administer nicotine if it is introduced into the bloodstream rather than the VTA. If so, it could be concluded that activation of $\beta 2$ -containing nAChRs in the VTA is sufficient to make nicotine rewarding, and drugs selectively targeting

such receptors might be useful in reducing nicotine addiction. Recently, another nAChR subunit, $\alpha 4$, has also been implicated in nicotine-induced reward⁶, so one would predict that reintroduction of the $\alpha 4$ subunit into the VTA of an $\alpha 2^{-/-}$ mouse would also reinstate nicotine self-administration behaviour. If so, $\alpha 4\beta 2$ receptors in the VTA are the key nAChRs necessary for nicotine addiction.

To explore the function of the VTA cells further, the authors examined the effects of the $\beta 2$ subunit on exploratory behaviour (in the absence of nicotine). Brain circuits linked to the VTA are involved in the development of adaptive responses to environmental stimuli, and this can be analysed by measuring exploratory behaviour and navigation, the difference being whether the animals investigate their surroundings as they move, or whether they travel through the environment without much interaction with it. The authors found that mice lacking $\beta 2$ showed increased navigation and decreased exploratory behaviour, implicating acetylcholine in these behaviours. Reintroduction of the $\beta 2$ gene into the VTA of these animals restored exploratory behaviour,

but did not affect navigation movements. This is a strong indication that endogenous acetylcholine triggers exploratory behaviour by binding to nAChRs on cells originating in the VTA.

Changeux and colleagues' experiments firmly connect exploratory behaviour with VTA cell function, as well as providing a causal link between a specific nAChR subunit and this behaviour. It remains to be determined which human behaviours are analogous to exploratory behaviour in the mouse. Might there be a link between exploratory behaviour, or risk-taking behaviours in general, and addictive drug self-administration? ■

Julie A. Kauer is in the Departments of Molecular Pharmacology, Physiology and Biotechnology, and of Neuroscience, Brown University, Box GB-4, Providence, Rhode Island 02912, USA. e-mail: Julie_Kauer@brown.edu

1. Maskos, U. *et al. Nature* **436**, 103–107 (2005).
2. Peto, R. *et al. Br. Med. Bull.* **52**, 12–21 (1996).
3. Champiaux, N. & Changeux, J. P. *Prog. Brain Res.* **145**, 235–251 (2004).
4. Picciotto, M. R. *et al. Nature* **374**, 65–67 (1995).
5. Picciotto, M. R. *et al. Nature* **391**, 173–177 (1998).
6. Tapper, A. R. *et al. Science* **306**, 1029–1032 (2004).

SOLID-STATE PHYSICS

Doping the undopable

Giulia Galli

Impurities that increase the number of electron carriers are essential in most bulk semiconductors. Introducing such foreign atoms into semiconductor nanocrystals is fiddly, and requires exact knowledge of the material's surface.

Almost a hundred years after the construction of the first 'bulk' (macroscopic) semiconductor device, Erwin *et al.* (page 91 of this issue)¹ present a mechanism to control the inclusion of transition-metal impurities in semiconductor nanocrystals — impurity inclusion is the process known as doping. This advance could allow the electronic and optical properties of nanocrystals to be engineered for applications ranging from solar cells to electronic devices that function using electron spin, rather than electric charge. An impurity introduced through doping could, for example, be used to inject a localized spin into one nanocrystal in an array, its interaction with other spin carriers forming the basis of a 'spintronic' device.

The physical effect that underlies the work of Erwin *et al.* is known as quantum confinement — the quantization, or splitting, at the nanoscale, of the continuum of electronic energy states present in a bulk crystal, such that the energy levels of semiconductor nanocrystals resemble those of giant molecules. This effect was discovered more than 20 years ago^{2,3}, almost simultaneously by groups in the United States and Russia working respectively on lead sulphide and cadmium

sulphide. These materials — compounds of elements from groups II and VI of the periodic table — are very similar to the crystals of galena, a naturally occurring form of lead sulphide, that Ferdinand Braun used in 1907 to build the first solid-state rectifier, ushering in the era of bulk semiconductor devices.

Bulk semiconductors are ubiquitous in device applications, because their properties may change when the number of active electrons (those that are free to move within the material and contribute to conduction) is modified — for example by doping with external impurities. As a result of quantum confinement, semiconductor nanocrystals not only possess markedly different optical properties from those of the bulk material, but they can also become extremely sensitive to doping. Exploiting this sensitivity can allow their physical and chemical properties to be controlled with atomic-scale precision, and can result in materials tailored to possess specific properties. Producing new materials 'atom by atom' is a revolution anticipated by Richard Feynman almost 50 years ago⁴ that is still in the making and represents a highly active field of interdisciplinary research.

Despite decades of experience in doping bulk

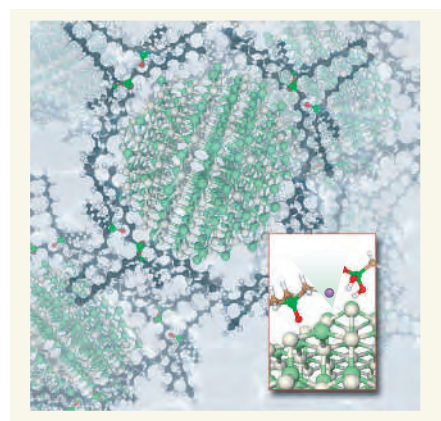


Figure 1 | Nanoscale doping. Semiconductor nanocrystals such as those investigated by Erwin *et al.*¹ can be engineered at the microscopic scale by the incorporation of impurities (doping). The main image is a ball-and-stick representation of cadmium selenide nanoparticles immersed in solution; the inset shows details of the surface structure interacting with model surfactants and with an impurity (purple sphere). Erwin *et al.* show that the incorporation of the impurity in a nanocrystal during growth is possible, but depends on the strength of its binding with surface atoms.

semiconductors — to build transistors, for example — doping nanocrystals has proved difficult. One explanation for this is the possible existence of intrinsic self-purification processes that could hamper the introduction of defects at the nanoscale. Also, depending on the preparation conditions, II–VI nanocrystals doped with transition metals may suffer from low crystallinity — that is, irregularities in their lattice structure. Nevertheless, there has been significant progress in recent years in doping II–VI nanocrystal solids and free-standing clusters^{5,6}.

Erwin *et al.*¹ suggest that some of the difficulties encountered in nanodoping are due to the fact that the mechanisms of impurity incorporation in bulk materials and at the nanoscale are profoundly different. At the macroscopic scale, thermodynamics provides the fundamental constraint on the amount of one solid that may be incorporated into another, yet the degree of doping achieved so far at the nanoscale is much lower than the thermodynamic limit. Thus, thermodynamic considerations seem to be irrelevant to impurity incorporation at the nanoscale. Rather, say Erwin *et al.*, it is kinetics that plays a key role — in particular, surface kinetics.

According to their model, an impurity present when the nanocrystal is synthesized can find its way in only if it can bind to the nanocrystal surface for a comparable time to that required for the crystal to grow in solution. The ability to dope and so modify a nanocrystal does not therefore stem from the equilibrium thermal diffusion of the 'guest' atom, as in a bulk solid, but rather from the binding energy of the guest atom to specific surface facets. In turn, the strength of this

binding depends on the morphology of the nanocrystal surface and on the surfactants — molecules that are present in the chemical solution in which the nanocrystal is synthesized and which may bind to or interact with the nanocrystal surface.

Confirmation that, at least in the case of II–VI nanocrystals, the surface binding energy is indeed the protagonist in the incorporation of impurities comes from a specific experiment¹ that nicely shows the progress made in the field of nanoscale manipulation. Using an appropriate core seed, Erwin *et al.*¹ grew a cadmium selenide (CdSe) shell with the desired lattice structure (Fig. 1) — a cubic lattice with a zinc blende structure, rather than the hexagonal lattice of the more usually adopted wurtzite structure. This CdSe shell had the surface morphology to which, according to calculations, an impurity of the transition metal manganese would best stick. In this way, the authors managed to use manganese to dope a previously undopable CdSe nanocrystal.

Binding energies between the nanocrystal and surfactants have also been found⁷ to play a key role in determining the shape of CdSe nanostructures, in particular whether they are rods or spheres. Defining the relationship between the microscopic structure and composition of a semiconductor nanocrystal and its function requires complex analysis. For this, *ab initio* simulations such as those on which Erwin and colleagues' experiment was based can prove most useful.

Surface morphology, structure and kinetics — identified by Erwin *et al.* as crucial to the doping of nanocrystals — are dominant in many other nanoscale phenomena. Examples are phase transformations⁸, the optical absorption and emission of group IV nanostructures, and the field of nanomechanics. This highlights some of the challenges of nanoscience research, where 'every atom counts'. At the nanoscale, details of the atomic structure (such as the surface structure) are often important, there are no known scalable models, and one must resort to the basic equations of quantum mechanics to investigate nanostructures. In addition, many of the processes occurring at the nanoscale are not in thermodynamic equilibrium, and thus simple thermodynamic considerations do not apply. Now we have a demonstration that, at least in some cases, these challenging problems are tractable. ■

Giulia Galli is at the Lawrence Livermore National Laboratory, PO Box 808, Livermore, California 94551, USA.

e-mail: galligygi1@llnl.gov

CANCER BIOLOGY

The weakest link?

Glenn Merlino

Cellular lineages are defined by master regulatory proteins that dictate their fate and ensure their survival. The dependence on such factors of tumours that are resistant to treatment may prove to be their Achilles' heel.

The pigment-producing cells in the skin — melanocytes — have a master regulator called MITF (for 'microphthalmia-associated transcription factor'). This factor is required for committing immature cells to the melanocyte lineage during development and is intimately involved in decisions regarding cell survival, growth and specialization (differentiation). Intuitively, one might expect that MITF would fiercely maintain melanocyte integrity, and discourage any deviation towards uncontrolled growth and malignancy. However, in this issue Garraway and colleagues (page 117)¹ report that melanoma cells tend to have extra, or 'amplified', copies of the gene that encodes MITF, and that under certain circumstances this gene can transform human melanocytes into cancerous cells. The melanoma cells still require MITF for survival, however, and for their characteristic resistance to drugs, presenting an unexpected target for the development of future therapies.

Garraway *et al.*¹ began by looking for alterations in the genomes of cell lines that make up a standard sample set called the NCI60 panel, which contains eight melanomas. Remarkably, although there had been no previous evidence that MITF is mutated in human cancer, the authors found that the chromosomal region containing the MITF gene (designated 3p13–3p14) was amplified in most of the melanoma cell lines. Expanding their analysis to include human tissue samples revealed that the MITF gene was also amplified (ranging from 5 to 119 copies) in about 10% of primary melanomas and up to 20% of metastatic melanomas, but not in moles (melanocytic nevi), which are considered a pre-malignant stage of some melanomas. Moreover, the amplification of MITF was significantly associated with decreased five-year survival in patients with metastatic melanoma.

MITF is an intriguing candidate for an amplified oncogene (a cancer-promoting gene), as there is compelling evidence that, in addition to its role in differentiation, it represses cell proliferation by activating the expression of inhibitors of the cell cycle^{2–5}. One of these, p16^{INK4a}, is a well-known melanoma tumour suppressor. How can a gene whose normal product restricts cell proliferation be amplified in growing tumours? One possible mechanism is through ordered alterations that uncouple MITF from proliferation, and perhaps also from differentiation. For example, it is likely that loss of p16^{INK4a} (or a mutation that

produces an equivalent effect) is a crucial early step in melanoma progression, and a prerequisite to MITF amplification (Fig. 1, overleaf).

In fact, the authors go on to show that in human melanocytes that have been genetically modified so that, among other things, p16^{INK4a} activity is blocked, MITF can transform the cells. However, this transforming activity only occurs when MITF is overexpressed in the presence of a mutated form of the BRAF protein, a vital signalling factor in melanocytes. This finding is significant, because BRAF mutations occur early in melanoma and are found in most nevi and melanomas^{4,5}.

What advantage, then, does enhanced MITF activity, whether through amplification of its gene or another mechanism, give the aspiring melanoma cell? The contributions of MITF the oncogene are undoubtedly as complex as those of MITF the master regulator. But clues may be gleaned from the actions of its targets, notably Bcl-2, a factor that promotes cell survival⁶. Because they must normally endure damaging ultraviolet radiation as well as the toxicity associated with biosynthesis of the melanin pigment, cells of the melanocyte lineage are primed for enhanced survival and depend heavily on factors that thwart cell-death pathways.

Lineage-specific survival mechanisms associated with MITF may account, at least in part, for the drug resistance that characterizes melanoma. Indeed, analysis of the available NCI60 pharmacological data revealed a significant correlation between MITF copy number and chemoresistance. Furthermore, Garraway *et al.* found that inhibiting MITF activity in melanoma cells harbouring extra copies of the MITF gene sensitized the cells to the growth-inhibitory effects of cisplatin and docetaxel — drugs currently used to treat melanoma, albeit relatively ineffectively. Agents that target MITF, or molecules further down the activation pathway that could be more suitable drug targets, may therefore enhance the therapeutic efficacy of conventional melanoma chemotherapy. It may also prove useful to screen for the presence of MITF amplifications before selecting treatment.

The discovery of MITF amplification in melanoma also backs up the theory of a link between cancer and stem cells — the immature cells that continuously divide to produce more highly specialized progeny. Melanocyte stem cells reside in the hair follicle⁷, where MITF has been implicated in their self-renewal and,

1. Erwin, S. C. *et al.* *Nature* **436**, 91–94 (2005).
2. Rossetti, R. & Brus, L. *J. Phys. Chem.* **86**, 172–177 (1982).
3. Efros, A. I. *Fiz. Tekh. Poluprovodn.* **16**, 1209–1214 (1982); *Sov. Phys. Semicond.* **16**, 772–775 (1982).
4. Feynman, R. www.zyvex.com/nanotech/feynman.html
5. Yu, D., Philippe, W. & Guyot-Sionnest, P. *Science* **300**, 1277–1280 (2003).
6. Raola, O. E. & Strouse, G. F. *Nano Lett.* **2**, 1443–1447 (2002).
7. Puzder, A. *et al.* *Nano Lett.* **4**, 2361–2365 (2004).
8. Huang, F. & Banfield, J. F. *J. Am. Chem. Soc.* **127**, 4523–4529 (2005).

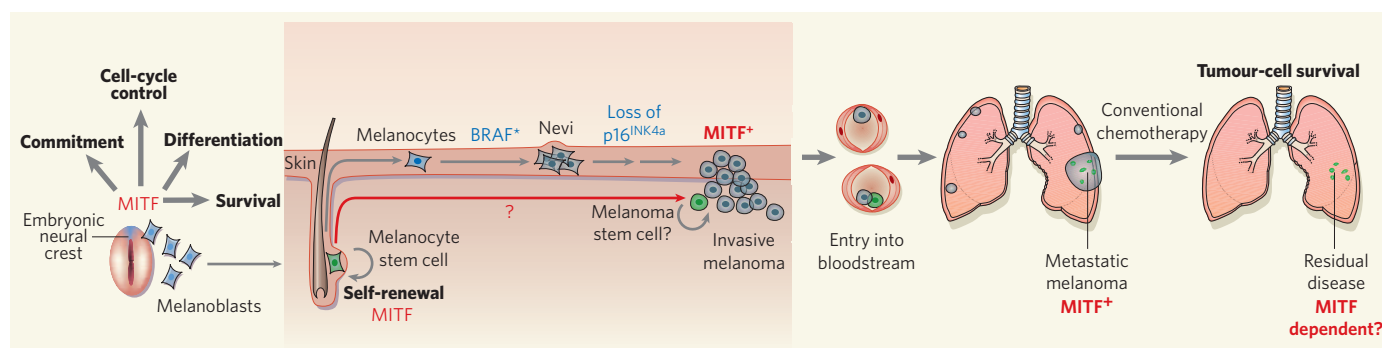


Figure 1 | A master regulator in melanoma. The melanocyte-specific microphthalmia-associated transcription factor (MITF) is required for the commitment of embryonic cells to the melanocyte lineage during development and controls melanocyte survival, growth and specialization. Processes shown in bold type require MITF. Committed melanocyte precursors (melanoblasts) migrate to skin, where they can differentiate into pigment-producing melanocytes, or remain as melanocyte stem cells (green) in the hair follicle. Downstream targets of MITF include differentiation factors, survival factors and cell-cycle inhibitors such as p16^{INK4a}. In melanoma, melanocytes progress through a series of

pathological stages that can include pre-malignant nevi (moles) and invasive melanoma before becoming metastatic melanoma. These stages have been associated with early mutations of BRAF (asterisk), loss of p16^{INK4a}, and other alterations. Garraway *et al.*¹ show that extra copies of the MITF gene (MITF⁺) can occur in primary and metastatic lesions. The presence of amplified MITF seems to promote resistance to conventional melanoma chemotherapy and tumour-cell survival. Surviving residual tumour tissue could be heavily populated by melanoma stem cells (shown in green, currently hypothetical), arising from normal stem cells or from more highly differentiated melanocytic cells (red arrow).

through its target Bcl-2, in survival⁸. Cancer stem cells, possessing the proliferative potential and self-renewal capacity of normal stem cells, occur in malignancies of the blood and in some solid tumours⁹, although they have not been seen in melanoma. Melanoma stem cells retaining the properties of melanocyte stem cells, but carrying amplified MITF and/or other key mutations, might stand a better chance of evading conventional chemotherapy by surviving as a dormant residual disease and recurring as lethal metastatic melanoma (Fig. 1).

But is MITF a double-edged sword? The amplification of MITF suggests that melanocyte stem cells' dependence on this factor for self-renewal and survival might be maintained, and even amplified, in their malignant counterparts. Over-reliance on lineage survival factors such as MITF could be a weak link in an otherwise unbreakable chain, providing an opportunity well worth exploiting therapeutically.

Glenn Merlino is in the Laboratory of Cell Regulation and Carcinogenesis, National Cancer

Institute, Bethesda, Maryland 20892-4264, USA.
e-mail: gmerlino@helix.nih.gov

1. Garraway, L. A. *et al.* *Nature* **436**, 117–122 (2005).
2. Loercher, A. E., Tank, E. M., Delston, R. B. & Harbour, J. W. *J. Cell Biol.* **168**, 35–40 (2005).
3. Carreira, S. *et al.* *Nature* **433**, 764–769 (2005).
4. Davies, H. *et al.* *Nature* **417**, 949–954 (2002).
5. Pollock, P. M. *et al.* *Nature Genet.* **33**, 19–20 (2003).
6. McGill, G. G. *et al.* *Cell* **109**, 707–718 (2002).
7. Nishimura, E. K. *et al.* *Nature* **416**, 854–860 (2002).
8. Nishimura, E. K., Granter, S. R. & Fisher, D. E. *Science* **307**, 720–724 (2005).
9. Dean, M., Fojo, T. & Bates, S. *Nature Rev. Cancer* **5**, 275–284 (2005).

CONSERVATION BIOLOGY

Where slugs may safely graze

Peter D. Moore

Grazing animals mow meadows to useful effect. From the results of experiments on newly established grassland, one such grazer, the little-considered slug, evidently has a big and beneficial influence on plant diversity.

Large mammalian grazers are allies of the conservation biologist. Sharp incisors, grinding molars and selectiveness in their choice of vegetation have made sheep, cattle and horses invaluable for grassland managers concerned with enhancing diversity or encouraging particular plant species.

But what of invertebrate grazers? Grasslands support large populations of small herbivores, from aphids to grasshoppers: what role do these lesser grazers play in the development of the plant composition of grassland habitats? This question lies at the heart of an experimental study, conducted by Buschmann and colleagues, and published in *Functional Ecology*¹, on the impact of slugs on the development of an area of sown grassland near Zurich in Switzerland.

Sir Arthur Tansley² recognized the significance

of grazing in determining the structure and composition of grasslands when he published his first survey of British vegetation types in 1911. His attention was mainly directed at sheep and rabbits, which then roamed in large numbers over the chalk hills of southern Britain. Such grazers remove a substantial proportion of grass production, but also, and perhaps even more importantly, graze selectively on preferred species. Invertebrate grazers might be expected to operate in a similar, if less apparent, fashion. Biomass removal, within certain limits, can have a positive effect on overall diversity by restricting the development of highly productive and aggressive plants, thus giving the smaller and slower-growing species an opportunity to germinate and become established³. The outcome of such selectivity depends on which members of

the plant assemblage prove most attractive to the grazer.

Buschmann and colleagues concentrated their attentions on slugs, in particular *Arion lusitanicus* (Fig. 1, overleaf), a large species (up to 10 cm in length) of western and southwestern Europe and a generalist herbivore. Even slugs have their dietary preferences, as any gardener can testify, and the seedling stages of plant development are often the most vulnerable to slug attack. So it is reasonable to suppose that slug impact on vegetation composition would be most profound in the early stages of plant colonization. In the more established stages, on the other hand, the effect of dietary selectivity may be less, but the possible effect of general biomass removal could still influence grassland structure. To test these hypotheses, the researchers set up an experimental



50 YEARS AGO

With the appearance of a new journal, *Virology* (pp. 140. New York: Academic Press, Inc.; 9 dollars per vol.), this useful, but ugly, word of doubtful parentage presumably takes its place as the official designation of the study of viruses.

From *Nature* 9 July 1955.

100 YEARS AGO

Even with things as they are, Oxford and Cambridge, though much injured by competitive examinations, have been far less injured than England in general; and this they owe to the residential system. Little thought of, perhaps neglected, by the builders, the head-stone of the educational edifice is here to be found. Where mind meets mind in the free intercourse of youth there springs from the contact some of the fire which, under our present system, is rarely to be obtained in any other way; and not only this, but many other priceless advantages in the battle for life are also conferred. To these influences we owe in large part all that is best in the English character, and so valuable are the qualities thus developed, or at least greatly strengthened, that we regard residential colleges as essential to the success and usefulness of the newer universities.

ALSO:

An Angler's Hours. By H. T. Sherringham. Mr. Sherringham deserves the thanks of all anglers who have an idle hour and no fishing for having re-published his essays in book form, and he who is forced by sad circumstance to enjoy his fishing vicariously will find his time well spent in our scribe's company... he despairs of nothing, but finds good in all; if there are no fish he can study nature, and if there is no water he can shrewdly meditate on the ways of fish and men; an hour with him and his rod by a troutless tarn is as good as an hour by the Kennet in the mayfly time... A word of praise is also due to the publishers, who have produced a book the size and print of which add to its convenience as an adjunct to a pipe, an easy chair, and idleness.

From *Nature* 6 July 1905.



Figure 1 | *Arion lusitanicus* — conservation agent.

grassland sown with rye grass (*Lolium perenne*) and white clover (*Trifolium repens*) on a former arable field that contained its own residual seed bank of weed and other plant species.

The surface soil was thoroughly mixed to avoid local patchiness in the seed bank, and a series of experimental 2 × 2-m plots was established, each surrounded by a slug-proof fence. Local slugs were placed in selected plots at a density of 22 individuals per plot during the first year, with an additional 10 slugs in subsequent years; this represents a high but realistic concentration of the molluscs. Wooden slug shacks provided shelter for these easily desiccated creatures in times of drought. The control plots were treated with molluscicide to prevent any inadvertent slug invasion. Analysis of the vegetation composition over the following three years provided the data needed to determine the effect of slug grazing.

In the first two years, the species richness and the diversity were lower in the slug-grazed plots than in controls. (Species richness is the number of species per plot; diversity also takes into account the proportions of different species, and is measured by the Shannon diversity index.) This result confirms the expectation that slug selection of seedlings

would reduce the number of species from the local seed bank that become established. In the third year of the experiment, however, species richness in the grazed plots was 23% higher than in the controls.

The reason for this enhancement of richness and diversity in the more mature stages can be attributed to the consistent removal of biomass by the slugs. The yield from primary productivity was reduced by around 25% as a result of slug grazing (comparable to the removal of biomass by sheep in a grazed pasture⁴). Holding back the development of dominance by fast-growing species provided an opportunity for the germination and establishment of less-competitive species, including annual plants. In other words, slug grazing permits the establishment of plant species that might otherwise find it difficult to maintain populations in developing grassland. So, on this account at least, slugs are good for diversity.

Slugs will never act as sheep substitutes by creating a pastorally idyllic landscape and inspiring poets. But they could well be an answer to the conservationist's prayer — silently grazing beneath our feet, they provide an alternative way to mow a meadow. ■

Peter D. Moore is in the Division of Life Sciences, King's College London, Franklin-Wilkins Building, 150 Stamford Street, London SE1 9NH, UK.
e-mail: peter.moore@kcl.ac.uk

1. Buschmann, H., Keller, M., Porret, N., Dietz, H. & Edwards, P. J. *Funct. Ecol.* **19**, 291–298 (2005).
2. Tansley, A. G. (ed.) *Types of British Vegetation* (Cambridge Univ. Press, 1911).
3. Grime, J. P. *Plant Strategies, Vegetation Processes, and Ecosystem Properties* (Wiley, Chichester, 2001).
4. Perkins, D. F. in *Production Ecology of British Moors and Montane Grasslands* (eds Heal, O. W. & Perkins, D. F.) 375–395 (Springer, Heidelberg, 1978).

NONLINEAR DYNAMICS

When instability makes sense

Peter Ashwin and Marc Timme

Mathematical models that use instabilities to describe changes of weather patterns or spacecraft trajectories are well established. Could such principles apply to the sense of smell, and to other aspects of neural computation?

Dynamical stability is ubiquitous in many systems — and more often than not is desirable. Travelling down a straight road, a cyclist with stable dynamics will continue in more or less a straight line despite a gust of wind or a bumpy surface. In recent years, however, unstable dynamics has been identified not only as being present in diverse processes, but even as being beneficial. A further exciting candidate for

this phenomenon is to be found in the realm of neuroscience — mathematical models^{1–3} now hint that instabilities might also be advantageous in representing and processing information in the brain.

A state of a system is dynamically stable when it responds to perturbations in a proportionate way. As long as the gust of wind is not too strong, our cyclist might wobble, but the

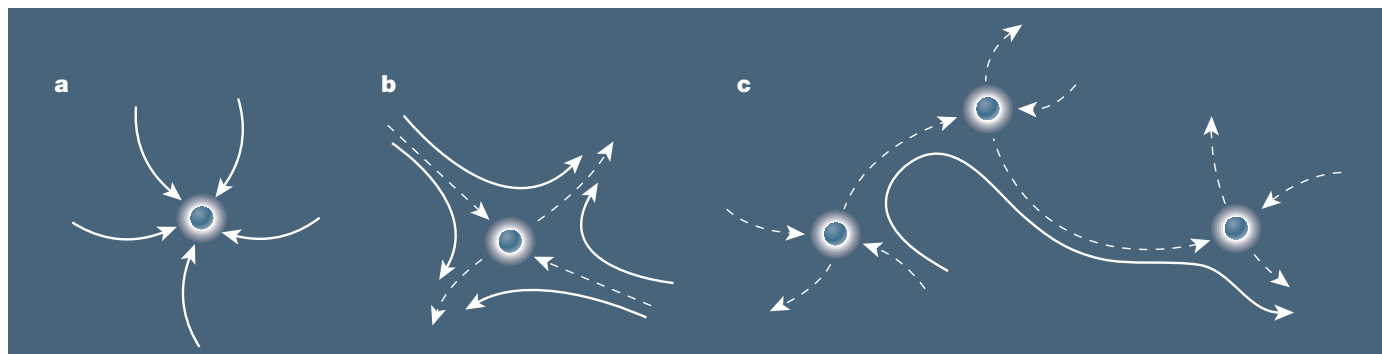


Figure 1 | Stable and unstable dynamics in 'state space'. **a**, A stable state with stationary dynamics. The system returns to the stable fixed point in response to small perturbations. **b**, An unstable saddle state is abandoned upon only small perturbations. The paths indicating possible evolutions of this system (solid lines) may pass close by such a state but will typically then move away. Only some of the exceptional

paths come back to the saddle state (dashed lines pointing inwards). **c**, A collection of saddles linked by 'heteroclinic' connections (dashed lines). The system evolves close to the heteroclinic connections between different saddles, lingering near one saddle state before moving on to the next. It is this last type of dynamics that several studies^{1–3,6,7} find in models of neural computation.

direction and speed of the cycle will soon return to their initial, stable-state values. This stable state can be depicted in 'state space' (the collection of all possible states of the system) as a sink — a state at which all possible nearby courses for dynamic evolution converge (Fig. 1a).

By contrast, at unstable states of a system, the effect of a small perturbation is out of all proportion to its size. A pendulum that is held upside-down, for example, although it can in theory stay in that position for ever, will in practice fall away from upright with even the smallest of disturbances. On a state-space diagram, this is depicted by paths representing possible evolutions of the system running away from the state, rather than towards it. If the unstable state is a 'saddle' (Fig. 1b), typical evolutions may linger nearby for some time and will then move away from that state. Only certain perturbations, in very specific directions, may behave as if the state was stable and return to it.

There is, however, nothing to stop the pendulum from coming back very close to upright if frictional losses are not too great. This is indicated on a state-space diagram by a path travelling close to what is known as a heteroclinic connection between two saddles. Heteroclinic connections between saddle states (Fig. 1c) occur in many different systems in nature. They have, for example, been implicated in rapid weather changes that occur after long periods of constant conditions⁴. Engineers planning interplanetary space missions⁵ routinely save enormous amounts of fuel by guiding spacecraft through the Solar System using orbits that connect saddle states where the gravitational pulls of celestial bodies balance out.

Several studies^{1–3,6,7} have raised the idea that this kind of dynamics along a sequence of saddles (Fig. 1c) could also be useful for processing information in neural systems. Many traditional models of neural computation share the spirit of a model⁸ devised by John Hopfield, where completion of a task is equivalent to the system becoming stationary at a stable state. Rabinovich *et al.*¹ and, more recently, Huerta

*et al.*² have shown that, in mathematical models of the sense of smell, switching among unstable saddle states — and not stable-state dynamics — may be responsible for the generation of characteristic patterns of neural activity, and thus information representation. In creating their models, they have been inspired by experimental findings in the olfactory systems of zebrafish and locusts⁹ that exhibit reproducible odour-dependent patterns.

Huerta *et al.*² model the dynamics in two neural structures known as the antennal lobe and the mushroom body. These form staging posts for processing the information provided by signals coming from sensory cells that are in turn activated by odour ingredients. Whereas activity in the mushroom body is modelled by standard means using stable dynamics, the dynamics of the antennal lobe is modelled in a non-standard way using networks that exhibit switching induced by instabilities. In these models, the dynamics of the neural system explores a sequence of states, generating a specific pattern of activity that represents one specific odour. The vast number of distinct switching sequences possible in such a system with instabilities could provide an efficient way of encoding a huge range of subtly different odours.

Both Rabinovich *et al.*¹ and Huerta *et al.*² interpret neural switching in terms of game theory: the neurons, they suggest, are playing a game that has no winner. Individual states are characterized by certain groups of neurons being more active than others; however, because each state is a saddle, and thus intrinsically unstable, no particular group of neurons can eventually gain all the activity and 'win the game'. The theoretical study¹ was restricted to very specific networks of coupled neurons, but Huerta and Rabinovich have now shown³ that switching along a sequence of saddles occurs naturally, even if neurons are less closely coupled, as is the case in a biological system.

Similar principles of encoding by switching along a sequence of saddles have also been investigated in more abstract mathematical

models (see refs 6, 7 for examples) that pinpoint possible mechanisms for directing the switching processes. One problem with these proposals from mathematical modelling^{1–3,6,7} is that there is no clear-cut experimental evidence of their validity in any real olfactory system. Nevertheless, all of the mathematical models rely on the same key features — saddles that are never reached but only visited in passing, inducing non-stationary switching — that have been shown to be relevant in other natural systems^{4,5}. In biology, the detection of odours by populations of neurons could be only one example.

Much remains to be done in fleshing out this view of natural processes in terms of dynamics exploiting saddle instabilities. Then we will see just how much sense instability really makes. ■

Peter Ashwin is at the School of Engineering, Computer Science and Mathematics, University of Exeter, Exeter, Devon EX4 4QE, UK.

Marc Timme is at the Max Planck Institute for Dynamics and Self-Organization, and the Bernstein Center for Computational Neuroscience, Bunsenstrasse 10, 37073 Göttingen, Germany.
e-mails: P.Ashwin@ex.ac.uk;
timme@chaos.gwdg.de

1. Rabinovich, M. *et al.* *Phys. Rev. Lett.* **87**, 068102 (2001).
2. Huerta, R. *et al.* *Neural Comput.* **16**, 1601–1640 (2004).
3. Huerta, R. & Rabinovich, M. *Phys. Rev. Lett.* **93**, 238104 (2004).
4. Stewart, I. *Nature* **422**, 571–573 (2003).
5. Taubes, G. *Science* **283**, 620–622 (1999).
6. Hansel, D., Mato, G. & Meunier, C. *Phys. Rev. E* **48**, 3470–3477 (1993).
7. Kori, H. & Kuramoto, Y. *Phys. Rev. E* **62**, 046214 (2001).
8. Hopfield, J. J. *Proc. Natl Acad. Sci. USA* **79**, 2554–2558 (1982).
9. Laurent, G. *Nature Rev. Neurosci.* **3**, 884–895 (2002).

CORRECTION

In the News and Views article "Granular matter: A tale of tails" by Martin van Hecke (*Nature* **435**, 1041–1042; 2005), an author's name was misspelt in reference 9. The correct reference is Torquato, S., Truskett, T. M. & Debenedetti, P. G. *Phys. Rev. Lett.* **84**, 2064–2067 (2000).

BRIEF COMMUNICATIONS

The vocal tract and the sound of a didgeridoo

Acoustic measurements show how a player can extract a range of timbres from this primitive instrument.

The Australian didgeridoo (or *yidaki* in the Yolngu language of northern Australia) is a simple musical instrument that, at the lips of an experienced player, is capable of a spectacular variety of timbres — considerably greater than those that can be coaxed from orchestral instruments, for example. To understand this phenomenon, we simultaneously measured the sound produced by the didgeridoo and the acoustic impedance of the player's vocal tract. We find that the maxima in the envelope of the sound spectrum are associated with minima in the impedance of the vocal tract, as measured just inside the lips. This acoustic effect is similar to the production of vowel sounds made during human speech or singing¹, although the mechanism is different, and leads to the surprising conclusion that experienced players are subconsciously using their glottis to accentuate the instrument's tonal variation.

The didgeridoo is traditionally made from a small tree trunk whose core has been eaten by termites. It usually plays only one note, with the player's lips vibrating at a frequency near a maximum in the instrument's acoustic impedance², which is a measure of how difficult it is to produce air vibration at a given frequency. Its musical value comes from the great variety of timbres and rhythms produced by experienced players as they alter their tongue position and mouth geometry.

Characteristic bands of emphasized frequencies, or formants, occur in the output sound and are similar to the higher formants of spoken vowels³. The acoustic mechanism responsible for the strong formant production has hitherto not been well understood because of the complexity of the interactions between the vocal tract, the lips and the instrument. The acoustic properties of the tract are particularly difficult to measure during performance because of the very high sound levels inside the mouth.

In these experiments, a broadband acoustical current was injected into the player's mouth, just inside the lips, using a tube (diameter, 3.7 mm). A probe microphone next to it (diameter, 1.5 mm) measured the pressure. The acoustic impedance was calculated as described^{4,5}.

Figure 1 shows an example of the vocal-tract impedance. The spectrum of the radiated sound was measured immediately after the impedance measurement, with the player maintaining a constant sound but with no injected broadband current, to avoid this

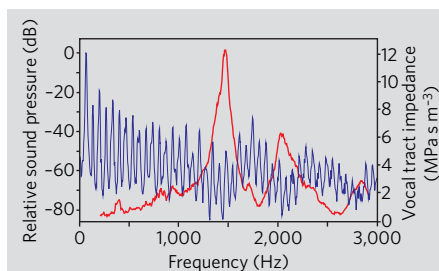


Figure 1 | Acoustic measurements from a didgeridoo performance. Spectrum of radiated sound (blue) and the magnitude of acoustic impedance of the vocal tract (red) measured just inside the lips of a didgeridoo player during performance. The player performs the 'high drone', produced with the tongue close to the hard palate, which generates a characteristic strong formant at 1.8 kHz. Similar measurements with the tongue in the low position revealed no strong impedance maxima and no strong formants. (For sound file and spectra⁸, see supplementary information.)

appearing as noise background in the sound spectrum. The sound output is reduced for frequencies at which the vocal-tract impedance is sufficiently large, and the acoustic flow through the lips is inhibited. For frequencies at which the vocal-tract impedance is small, flow into the instrument is greater and more sound is produced. Results were similar on other didgeridoos and with other players. (For details, see supplementary information.) A linear regression of the frequencies of formant maxima/minima against impedance minima/maxima has a slope of 0.93, with correlation coefficient of 0.98 (46 measurements from three players; results not shown).

These results indicate that strong formants in the sound depend on the presence of strong resonances in the tract, and this in general requires the glottis to be partially closed to enhance

reflection and to prevent the resonant high-frequency components being absorbed in the resistive impedance of the lungs. We conclude that a major difference between a novice and an experienced player is a learned, but usually subconscious, ability to reduce the glottal opening (a similar difference has been proposed for brass-instrument players⁶). The effect of tongue position on the pitch and timbre of orchestral lip-valve instruments is much more modest because of their relatively narrow bore, but is still musically significant⁷. An understanding of tract-lip-bore interactions in the didgeridoo, where the effects are pronounced, is likely to lead to an improved understanding of these more subtle effects in orchestral instruments.

Alex Tarnopolsky*, Neville Fletcher†‡, Lloyd Hollenberg‡, Benjamin Lange*, John Smith*, Joe Wolfe*

*School of Physics, University of New South Wales, Sydney, New South Wales 2052, Australia e-mail: j.wolfe@unsw.edu.au

†Research School of Physical Sciences, Australian National University, Canberra 0200, Australia

‡School of Physics, University of Melbourne, Parkville, Victoria 3010, Australia

1. Sundberg, J. *The Science of the Singing Voice* (Northern Illinois Univ., De Kalb, 1987).
2. Fletcher, N. H. & Rossing, T. D. *The Physics of Musical Instruments* 2nd edn (Springer, New York, 1998).
3. Fletcher, N. H. *Austr. Aboriginal Stud.* **1**, 28–37 (1983).
4. Epps, J., Smith, J. R. & Wolfe, J. *Meas. Sci. Technol.* **8**, 1112–1121 (1997).
5. Smith, J. R., Henrich, N. & Wolfe, J. *Proc. Inst. Acoust.* **19**, 315–320 (1997).
6. Mukai, S. *J. Otolaryngol. Jpn* **92**, 260–270 (1989).
7. Wolfe, J., Tarnopolsky, A. Z., Fletcher, N. H., Hollenberg, L. C. L. & Smith, J. *Proceedings of the Stockholm Music Acoustics Conference* (ed. Bresin, R.) 307–310 (2003).
8. Music Acoustics www.phys.unsw.edu.au/~jw/yidaki.html

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.
doi:10.1038/43639a

PALAEOCLIMATOLOGY

The record for marine isotopic stage 11

The marine isotopic stage 11 (MIS 11) is an extraordinarily long interglacial period in the Earth's history that occurred some 400,000 years ago and lasted for about 30,000 years. During this period there were weak, astronomically induced changes in the distribution

of solar energy reaching the Earth. The conditions of this orbital climate forcing are similar to those of today's interglacial period^{1,2}, and they rendered the climate susceptible to other forcing — for example, to changes in the level of atmospheric carbon dioxide. Here we use

ice-core data from the Antarctic Vostok core to reconstruct a complete atmospheric carbon dioxide record for MIS 11. The record indicates that values for carbon dioxide throughout the interglacial period were close to the Earth's pre-industrial levels and that both solar energy and carbon dioxide may have helped to make MIS 11 exceptionally long. Anomalies in the oceanic carbonate system recorded in marine sediments at the time³, for example while coral reefs were forming, apparently left no signature on atmospheric carbon dioxide concentrations.

During MIS 11 (ref. 3), there is evidence that large parts of the ocean were exceptionally

warm and changes in the oceanic carbonate system were drastic³. A complete MIS 11 carbon dioxide record has therefore been long-awaited. We revisited the Vostok ice-core data, correcting the ice stratigraphy for flow disturbances and establishing a common chronology with the record from Dome C of the European Project for Ice Coring in the Antarctic (EDC)⁴. The first 3,310 m of the Vostok core provide a climatic record back to MIS 11 ice, but below this there are indications that the record is disturbed, probably because of ice-flow anomalies^{5,6}.

The climatic record provides strong indications in favour of a depth reversal in the stratigraphic order of the layer, which corresponds to the 3,320–3,345 m interval (Fig. 1). Once the sequence is reversed, the carbon dioxide and deuterium phase relationship is in the right order. An important control for the validity of our Vostok reconstruction is the remarkable agreement with the carbon dioxide, methane and dust measurements from the EDC core, which highlight the changes that occurred during the MIS 11–12 transition (Fig. 1b).

How can the existence of an inverted MIS 11–12 transition at Vostok be explained? Overturned folds are often present in basal layers of ice sheets. Open folds associated with transient flow over bed topography can be assumed to exist upstream from Vostok, which is far from an ice divide. Horizontal simple shear would have turned such open folds into recumbent folds, and subsequent longitudinal extension could have initiated boudinage of the upper limb of the thin MIS 11–12 transition layer, explaining the disruption of the climatic sequence⁷. The absence of the upper limb of the fold in the Vostok core could also be explained by the normal faulting that is associated with intense shear in ice from the glacial period, which is related to the preferential orientation of the optical axes of ice crystals⁸.

To correct the environmental record for ice-flow perturbation, we assumed that the depth interval between 3,321 and 3,345 m was inverted, and a common dating with the EDC core was obtained (Fig. 1). The reconstructed Vostok carbon dioxide record plotted against the EDC timescale is shown in Fig. 1 for the period from 380,000 to 436,000 years before the present, which corresponds to the *in situ* depth interval of

3,220–3,350 m. The mean time resolution is 1,400 years. The carbon dioxide values (measurement accuracy, ± 3 p.p.m.) are those previously published^{5,9} and range between 289 and 266 p.p.m., with a mean value of 278 p.p.m., which is very close to the Earth's pre-industrial carbon dioxide level (280 p.p.m.).

There are several conclusions that can be drawn from our reconstructed Vostok MIS 11 carbon dioxide record. First, as the MIS 11 carbon dioxide concentrations are comparable to pre-industrial values for the duration of the interglacial period (about 30,000 yr), carbon dioxide forcing, together with solar energy, probably kept MIS 11 exceptionally long^{2,6}. Second, the MIS 11 carbon dioxide concentrations were not especially high; for instance, they are lower than during stage 9 (ref. 5). This has to be taken into account in deciding whether MIS 11 was a particularly warm interglacial period and whether or not the oceanic carbonate system was so unusual³. Third, unlike the three following interglacials (MIS 5, 7 and 9) and the Holocene epoch⁵, MIS 11 does not show an early carbon dioxide peak; we therefore cannot consider the highest carbon dioxide values at the start of these periods as being the natural mode of interglacials¹⁰. Finally, as for MIS 5 (refs 5, 11), carbon dioxide concentrations show interglacial values for several thousands more years after the start of the Antarctic cooling that led to the next ice age.

Dominique Raynaud*, **Jean-Marc Barnola***,
Roland Souchez†, **Reginald Lorrain‡**,
Jean-Robert Petit*, **Paul Duval***,
Vladimir Y. Lipenkov‡

*Laboratoire de Glaciologie et
Géophysique de l'Environnement, CNRS,
38402 Saint-Martin-d'Hères, France
e-mail: raynaud@lgge.obs.ujf-grenoble.fr

†Laboratoire de Glaciologie, Université Libre de
Bruxelles, Brussels 1050, Belgium

‡Arctic and Antarctic Research Institute,
St Petersburg 199397, Russia

- Berger, A. & Loutre, M. F. in *Earth's Climate and Orbital Eccentricity: The Marine Isotope Stage 11 Question* (eds Droxler, A. W., Poore, R. Z. & Burckle, L. H.) 17–26 (Am. Geophys. Union, Washington DC, 2003).
- Berger, A., Loutre, M. F. & Crucifix, M. *Surv. Geophys.* **24**, 117–138 (2003).
- Droxler, A. W., Poore, R. Z. & Burckle, L. H. (eds) *Earth's Climate and Orbital Eccentricity: The Marine Isotope Stage 11 Question* (Am. Geophys. Union, Washington DC, 2003).
- EPICA community members *Nature* **429**, 623–628 (2004).
- Petit, J. R. et al. *Nature* **399**, 429–436 (1999).
- Raynaud, D. et al. in *Earth's Climate and Orbital Eccentricity: The Marine Isotope Stage 11 Question* (eds Droxler, A. W., Poore, R. Z. & Burckle, L. H.) 27–40 (Am. Geophys. Union, Washington DC, 2003).
- Cunningham, J. & Waddington, E. D. *J. Glaciol.* **36**, 269–272 (1990).
- Ramsay, J. G. *Folding and Fracturing of Rocks* (McGraw-Hill, New York, 1967).
- Pépin, L., Raynaud, D., Barnola, J.-M. & Loutre, M. F. *J. Geophys. Res.* **106**, 31885–31892 (2001).
- Ruddiman, W. F. *Clim. Change* **61**, 261–293 (2003).
- Cuffey, K. M. & Vimeux, F. *Nature* **412**, 523–527 (2002).
- Goujon, C., Barnola, J.-M. & Ritz, C. *J. Geophys. Res.* **108**, doi:10.1029/2002JD003319 (2003).

Competing financial interests: declared none.
doi:10.1038/43639b

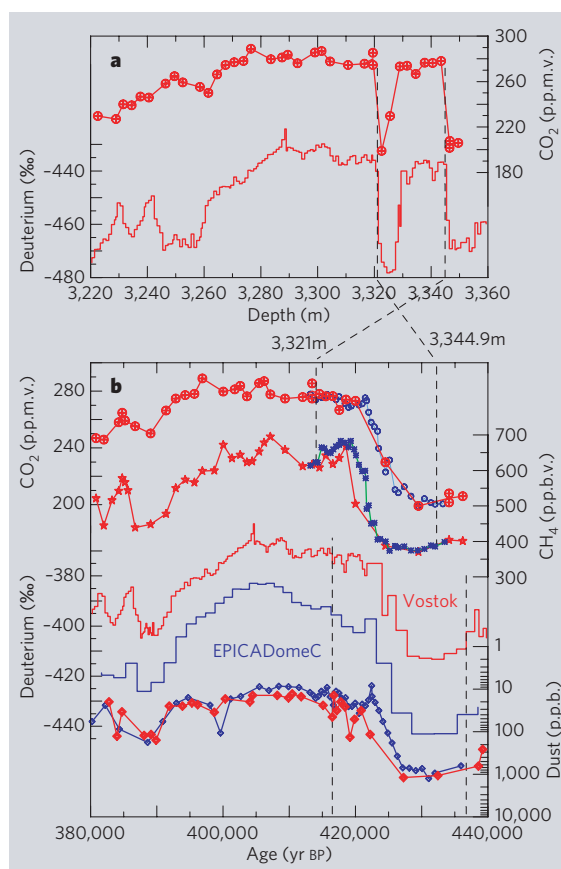


Figure 1 | The Vostok deuterium, carbon dioxide, methane and dust records of MIS 11. **a**, Original deuterium (red continuous line) and carbon dioxide (red circles) records versus depth. **b**, Corrected records for stratigraphic disturbances. To obtain a common chronology, the Vostok deuterium record has been fitted with the one from EPICA Dome C (EDC; blue continuous line) and plotted against the EDC timescale⁴. The Vostok carbon dioxide, methane and dust records (in red) are compared with the published EDC record⁴ (in blue), which currently covers the MIS 11–12 transition and the initial part of MIS 11. Dashed lines delimit the inverted parts of the Vostok record and correspond to the different ages for deuterium and carbon dioxide that reflect the gas age–ice age difference. This difference has been calculated with a firm densification model that includes heat diffusion¹². Temperatures and accumulation rates in the model were estimated from the deuterium–temperature and temperature–accumulation relationships used for establishing the Vostok GT4 chronology⁵ in this depth range. The Vostok deuterium record has been shifted by +65‰ to aid visual comparison between different curves. J. Chappellaz, E. Monin and U. Siegenthaler contributed some of these data.

NEWS & VIEWS MARS

Twin studies on Mars

David C. Catling

The twin Mars Exploration Rovers don't themselves range widely, but the observations they make do. Information on partial solar eclipses, salty rocks and magnetic dust are among the latest highlights of the rovers' findings.

Sending probes to Mars is a risky business. Roughly half the spacecraft ever sent have crashed, burnt up or simply missed the planet altogether. In contrast, Spirit and Opportunity, NASA's Mars Exploration Rovers, landed successfully on opposite sides of Mars in January 2004, and continue operations today. Six papers^{1–6} in this issue announce new analyses of the scientific data being beamed back to Earth.

Before the rovers landed, an array of remote-sensing data was already available, mostly from NASA's recent orbiters sent to Mars. Golombek *et al.*¹ assess the accuracy of that information, and conclude that physical characteristics of the martian surface — such as rock abundance, dustiness and topography — were accurately forecast beforehand. The prediction for average atmospheric density was 8% too high, so parachutes opened late. Fortunately, both spacecraft still descended safely, which for Spirit was critically aided by small rockets that compensated for winds. Nonetheless, when you look at Mars as a small red speck in the night sky, it seems an incredible feat to land spacecraft on specific parts of it.

The rovers have been making their own celestial observations, picturing silhouettes of Mars' moons as they move across the disk of the Sun². Viewed from the martian tropics, Phobos, the larger moon, rises in the west and scoots across the sky in several hours in a low orbit. In a transit, Phobos covers about a

quarter of the solar disk. Deimos, which is smaller and with a much higher orbit, appears only as a dot. Analysis of six transits implies that the moons' positions are within the 10–20-km uncertainty range of astronomical predictions, but further transit observations could improve orbital models. Because of the solid-body tides that it raises on Mars, Phobos is losing height and will strike Mars within 40 million years, assuming it remains intact⁷. Meanwhile, Deimos is drifting away. Greater accuracy in the acceleration rates of both moons would provide better information on the moons' futures and histories, as well as on Mars' interior (from tidal physics).

In contrast with the predictions for moon transits and landing safety, geological predictions have proved less dependable. Spirit landed inside a feature known as Gusev crater (Fig. 1), the aim being to find a dried-up crater lake. But Gusev's interior has turned out to be largely volcanic, strewn with wind-blown dust and basaltic rocks, created by volcanic activity, that have been ejected by impacts. Opportunity's site, Meridiani Planum, was correctly predicted to contain abundant haematite (Fe_2O_3 ; ref. 8). But no one imagined that the haematite would exist as blueberry-sized concretions (nodules precipitated from groundwater), or that the concretions would originate within sulphate-rich sedimentary rocks⁹.

Early in the mission, one scientist on the Spirit team complained that he felt stuck in a

“basalt prison” when he saw the riches available to Opportunity. Haskin *et al.*³ now report that rocks at the Spirit site are coated in material enriched in sulphur, oxidized iron, chlorine and bromine (Fig. 2a). The oxidized iron in the rocks has apparently been extracted from the basaltic mineral olivine [$(\text{Mg}, \text{Fe})_2\text{SiO}_4$]; similarly, oxidized iron in soils has been derived from soil olivine⁴. Salts in the subsurface soil seem to have been separated according to solubility, and the filled cavities and veins of rock interiors are enriched in bromine, an element known to form highly soluble salts.

All these observations could be explained by interaction with acidic water and subsequent evaporation. But only small quantities of transient water are required, not the pools of water thought to have produced sulphate-rich sediments at Meridiani⁹. Mars undergoes ice ages that are more extreme than those on Earth: ice migrates from the poles to the tropics as the martian spin axis periodically tips over¹⁰. Ice deposited in the tropics could form films of low-temperature brines that chemically alter basaltic rocks. Indeed, Yen *et al.*⁴ argue this case for the soils generally. At the same time, pristine olivine indicates that neither soils nor dust have ever been soaked in water for long periods^{4,5}.

Where did the soil come from? Some components of the soil (Fig. 2b), such as the haematite at Meridiani, are remnants of sedimentary deposits. And about 1% of soil

NASA/JPL/CORNELL (J. BEIL)

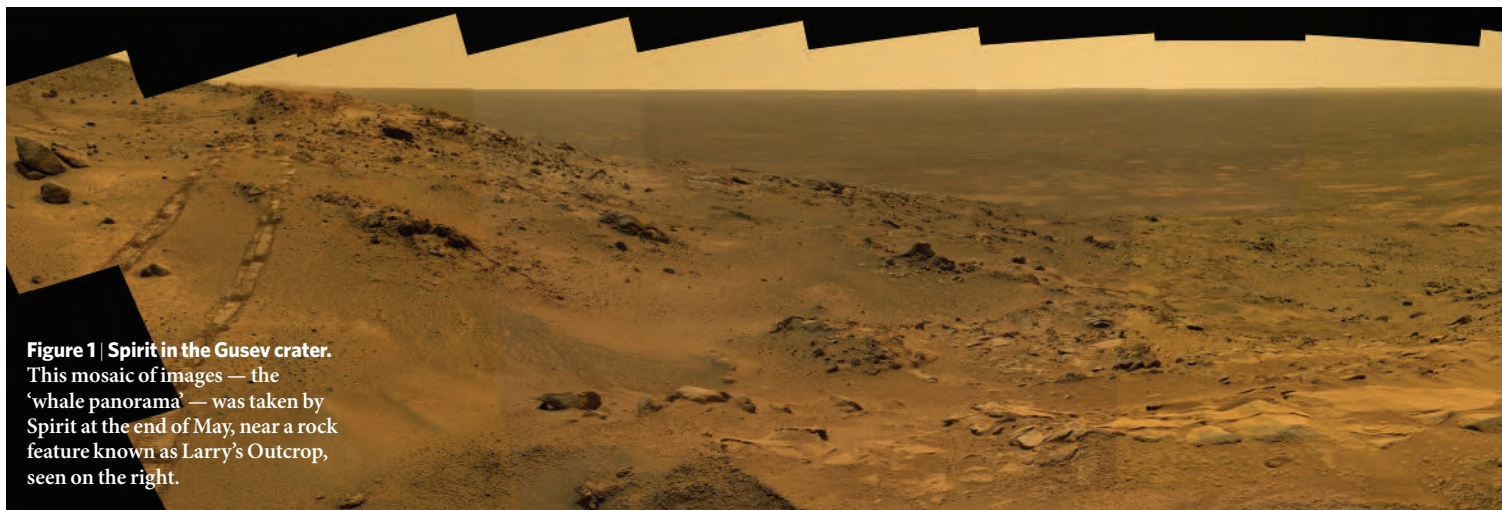


Figure 1 | Spirit in the Gusev crater. This mosaic of images — the ‘whale panorama’ — was taken by Spirit at the end of May, near a rock feature known as Larry’s Outcrop, seen on the right.

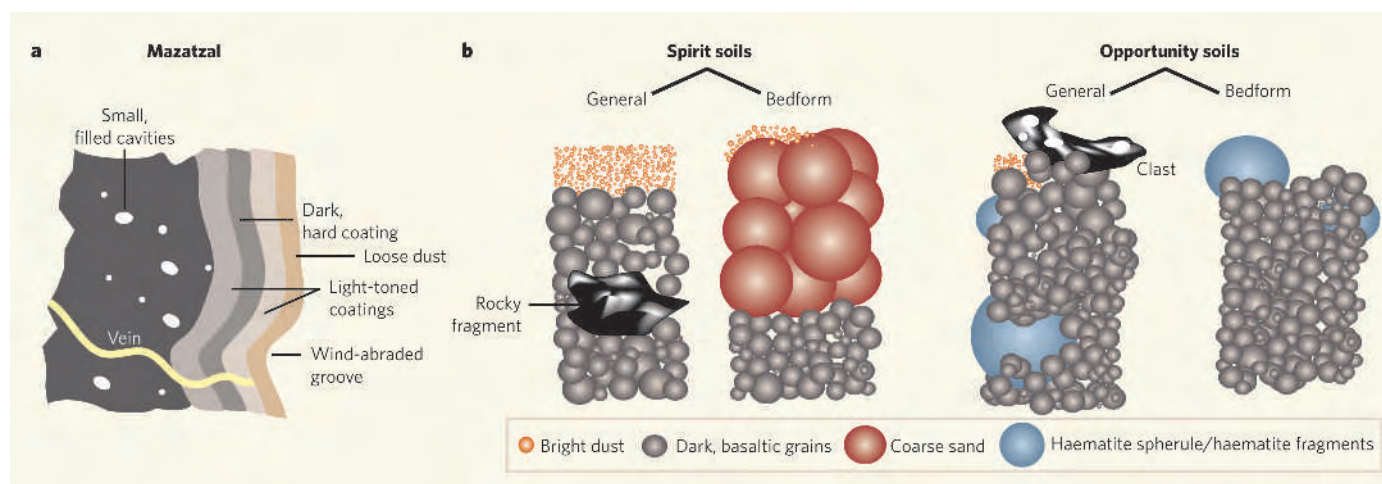


Figure 2 | Rocks and sands of Mars. **a**, One basaltic rock, Mazatzal, at the site of the Spirit rover, contains features that indicate interaction with water. Light-toned and dark coatings are enriched in sulphur, chlorine and oxidized iron, which can be explained by interaction with acidic water and subsequent evaporation. (Coating thicknesses are exaggerated, colours in both **a** and **b** are arbitrary.) **b**, Soils at each rover site have four components⁴. At Spirit's site, generally bright dust at the surface is underlain by dark

basaltic soil (with grain sizes up to about 100 μm), and interspersed rocky fragments. Bedforms, such as ripples, have coarse millimetre-sized sand at their surface which contains magnetite. At Opportunity's site, bright dust is found in patches, but otherwise the soil is made up of haematite spherules or their fragments, and dark basaltic grains. Clasts (bits broken from larger rocks) are also occasionally found near the surface, often with vesicles, which are holes that formed when gas was released from the parent lava.

arguably came from meteorites, resulting in a high abundance of nickel⁴. Wind has supplied both rover sites with the major components of bright dust and dark basaltic sand⁴.

How the wind interacts with the surface is a topic examined by Sullivan *et al.*⁶. It is difficult to measure martian atmospheric circulation, but dark and bright streaks downwind from craters indicate wind direction. Seen from orbit, the crater where Opportunity landed had a bright streak to its southeast. Surface observations confirm that the streak derives from bright dust-sized particles, probably deposited from the air during a dust storm. Dunes and ripples on Mars have been predicted to consist of much coarser particles, 200 μm or more in size, that are driven along the surface — with smaller particles being lofted into suspension by wind turbulence¹¹. However, some ripples at Meridiani, inside craters and pits, contain surface sand well below the predicted threshold size. Theory therefore needs revising: near-surface wind

turbulence is less effective at suspending particles than predicted⁶.

A nearly uniform elemental composition shows that the bright dust is distributed globally by winds. But the source of the dust's magnetism has remained elusive. On the basis of compositional analysis of dust captured on magnets, magnetite (Fe_3O_4) is now identified as responsible⁵. Evidently, the dust is a mixture of basalt and oxidized minerals, but the origin of the latter is unclear. Examining the dust morphology would help, but the rovers' microscope lacks the magnification to see micrometre-sized dust or its mineral components. Fortunately, NASA's Phoenix lander, due for launch in 2007, carries both a colour optical microscope and an atomic force microscope that will open up these unseen vistas.

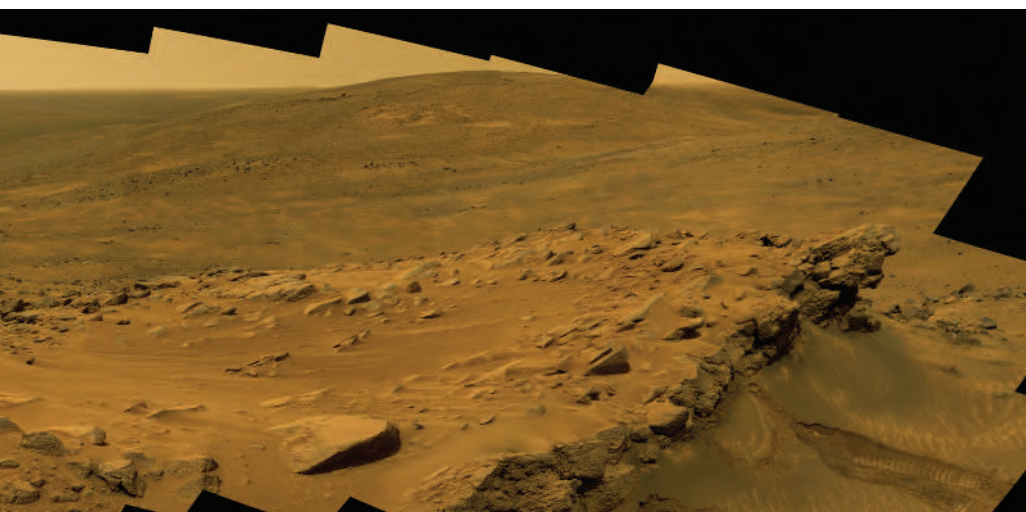
The overall picture is that much of the martian surface consists of volcanic soils and that Mars is sulphur-rich, with a geochemistry — like Mars' location — lying between that of

Earth and of Io, one of Jupiter's moons. Sulphur and oxidizing waters have reacted with the basaltic surface, producing sulphates, iron oxides, and presumably a clay or silica component that remains to be characterized.

Indeed, the largest gap in our understanding of the martian surface is caused by the absence of comprehensive data on mineral chemistry. So far, instruments have only measured elemental abundance or partial mineralogy from spectra. NASA's Mars Science Laboratory, slated for a 2009 or 2011 launch, will have X-ray diffraction capability¹². With definitive mineralogy, we should get answers to such tantalizing questions as how much of the salts come from evaporated seas, from meltwater at the end of ice ages, or — like ancient acid burns — from past reactions of volcanic gases with martian rocks and soils.

David C. Catling is in the Department of Atmospheric Sciences and the Astrobiology Program, University of Washington, Seattle, and the Department of Earth Sciences, University of Bristol, Queen's Road, Bristol BS8 1RJ, UK. e-mail: David.Catling@bristol.ac.uk

1. Golombek, M. P. *et al.* *Nature* **436**, 44–48 (2005).
2. Bell, J. F. III *et al.* *Nature* **436**, 55–57 (2005).
3. Haskin, L. A. *et al.* *Nature* **436**, 66–69 (2005).
4. Yen, A. S. *et al.* *Nature* **436**, 49–54 (2005).
5. Goetz, W. *et al.* *Nature* **436**, 62–65 (2005).
6. Sullivan, R. *et al.* *Nature* **436**, 58–61 (2005).
7. Burns, J. A. in *Mars* (eds Kieffer, J., Jakosky, B., Snyder, C. & Matthews, M. U.) 1283–1301 (Univ. Arizona Press, Tucson, 1992).
8. Christensen, P. R., Morris, R. V., Lane, M. D., Bandfield, J. L. & Malin, M. C. *J. Geophys. Res.* **106**, 23873–23885 (2001).
9. Squyres, S. W. *et al.* *Science* **306**, 1709–1714 (2004).
10. Haberle, R. M., Murphy, J. R. & Schaeffer, J. *Icarus* **161**, 66–89 (2003).
11. Greeley, R. & Iversen, J. D. *Wind as a Geological Process* (Cambridge Univ. Press, 1985).
12. Blake, D. F. *et al.* *Lunar Planet. Sci. Conf.* XXXVI, abstr. 1608 (2005).



ANALYSIS MARS

Assessment of Mars Exploration Rover landing site predictions

M. P. Golombek¹, R. E. Arvidson², J. F. Bell III³, P. R. Christensen⁴, J. A. Crisp¹, L. S. Crumpler⁵, B. L. Ehlmann⁶, R. L. Fergason⁴, J. A. Grant⁷, R. Greeley⁴, A. F. C. Haldemann¹, D. M. Kass¹, T. J. Parker¹, J. T. Schofield¹, S. W. Squyres³ & R. W. Zurek¹

Comprehensive analyses of remote sensing data during the three-year effort to select the Mars Exploration Rover landing sites at Gusev crater and at Meridiani Planum correctly predicted the atmospheric density profile during entry and descent and the safe and trafficable surfaces explored by the two rovers. The Gusev crater site was correctly predicted to be a low-relief surface that was less rocky than the Viking landing sites but comparably dusty. A dark, low-albedo, flat plain composed of basaltic sand and haematite with very few rocks was expected and found at Meridiani Planum. These results argue that future efforts to select safe landing sites based on existing and acquired remote sensing data will be successful. In contrast, geological interpretations of the sites based on remote sensing data were less certain and less successful, which emphasizes the inherent ambiguities in understanding surface geology from remotely sensed data and the uncertainty in predicting exactly what materials will be available for study at a landing site.

Selection of the Mars Exploration Rover (MER) landing sites took place over a three-year period in which engineering constraints were identified, of 155 possible sites two were selected, surface environments and safety considerations were developed, and the potential scientific knowledge to be obtained at the sites was considered¹. Landing sites in the Gusev crater and at Meridiani Planum were selected because they appeared acceptably safe for MER landing and roving and had strong morphologic or mineralogical indicators of having had liquid water in the past. The two sites therefore appeared capable of addressing the science objectives of the MER missions: to determine the aqueous, climatic and geologic history of sites on Mars where conditions may have been favourable to the preservation of evidence of possible pre-biotic or biotic processes.

Engineering constraints important to the selection included: latitude (10°N–15°S) for maximum solar power; elevation (<–1.3 km) for sufficient atmosphere to slow the descent of the lander; low horizontal winds, shear and turbulence in the last few kilometres to minimize horizontal velocity; low 10-m-scale slopes to reduce airbag spin-up and bounce; low to moderate rock abundance to reduce abrasion or stroke-out of the airbags; and a radar-reflective, load-bearing surface that is not dominated by fine-grained dust, and is thus safe for landing and roving¹. In selecting the MER landing sites these engineering constraints were addressed via comprehensive evaluation of surface and atmospheric characteristics from existing remote sensing data and models as well as targeted orbital information acquired from the Mars Global Surveyor (MGS) and Mars Odyssey.

This evaluation resulted in a number of predictions of the surface characteristics of the sites¹, which are tested in this paper. Relating remote sensing signatures to surface characteristics at landing sites allows these sites to be used as ground truth for the orbital data, is essential for selecting and validating landing sites for future missions, and is required for correctly interpreting the surfaces and materials globally present on Mars.

General predictions

General predictions of the surface characteristics made before landing were that both landing sites would be safe for the MER landing system and trafficable by the rovers¹. At Gusev crater, the available data suggested its appearance would be generally similar to the Viking Lander (VL) and Mars Pathfinder (MPF) landing sites, roughly as dusty but less rocky (Fig. 1). The geologic setting of the flat-floored Gusev crater at the end of Ma'adim Vallis, one of the largest branching valley networks on the planet, argued strongly that the materials inside were deposited in a crater lake^{2,3}. The Late Hesperian/Early Amazonian cratered plains³ upon which the landing site was principally sited showed little to reveal their origin with volcanic, aeolian (wind-formed) and lacustrine (lake-deposited) sedimentary materials as possibilities. If the surface materials were not lacustrine, it was hoped that the impacts would provide access to deeper materials that were¹.

At Meridiani Planum, the available data suggested a low-albedo surface with few rocks and little dust that would look completely unlike any of the VL or MPF landing sites¹ (Fig. 2). Evaluation of the geologic setting of Meridiani suggested a flat to gently rolling plain composed of basaltic sand with haematite and sparse outcroppings of a thin bright layer^{4–6}. The identification of coarse-grained haematite in MGS Thermal Emission Spectrometer (TES) spectra and the geologic setting from Thermal Emission Imaging System (THEMIS) data argued for direct precipitation of haematite from highly oxygenated iron-rich lake waters, or via alteration by percolating fluids after burial^{5,6} although alternative explanations were also possible¹.

All of the predictions of the general physical characteristics of the surface appear correct in the exploration of the landing sites by the rovers. In addition, we have compared the specific remote sensing data at the same landing and traverse locations^{7,8} to the surface characteristics observed by the rovers. The predictions of the materials that would be found scientifically at the two landing sites have proved less definitive.

¹Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA. ²Washington University, St Louis, Missouri 63130, USA. ³Cornell University, Ithaca, New York 14853, USA. ⁴Arizona State University, Tempe, Arizona 85287, USA. ⁵New Mexico Museum of Natural History and Science, Albuquerque, New Mexico 87104, USA. ⁶Oxford University, OX1 3PG, UK; ⁷Smithsonian Institution, Washington DC 20560, USA.

Atmosphere

Although the atmosphere (density and winds, in particular) was a key concern for safely landing the MER rovers⁹, there was no instrumentation to measure the atmosphere encountered directly. An advisory team was assembled to assess the available information in a timely manner, especially for Spirit (Opportunity would land three weeks later). In the case of the density (pressure and temperature), the deceleration profile was useful to reconstruct the atmosphere (as was done for Pathfinder¹⁰). A temperature profile was obtained with simple assumptions from a density profile derived from the deceleration curve and aeroshell drag properties.

A preliminary reconstruction immediately after landing was within the one-standard-low-deviation uncertainty bounds of the a priori atmosphere model¹ through most of the descent (as adjusted for the December 2003 dust storm using MGS TES temperature profiles¹¹ just before landing) for both Spirit and Opportunity. The mean model temperatures were within ~ 5 K of the preliminary reconstructed profile throughout the atmosphere for Spirit, with the model being warm below ~ 15 km and cool between 20 and 35 km. The Opportunity model showed a similar pattern of differences, but the deviations were as large as ~ 15 K (although the reconstructed profile was more uncertain). Both models overestimated the mean densities by an average of 8% throughout the atmosphere owing to uncertainties below 5 km. The MGS Mars Orbiter Laser Altimeter (MOLA)¹² elevations that were used to construct the density profiles are in excellent agreement with the elevations determined via radio tracking (within 7 m for Spirit and <1 m for Opportunity), thereby providing an accurate reference for the atmospheric model⁹.

Determining the wind and wind shear that the flight system encountered during descent is extremely difficult (the response has to be separated from other effects and the intrinsic flight system behaviour). It appears that the winds encountered were within the expectations based on the modelling. In a qualitative sense it seems

that the Meridiani landing site was less windy than at the Gusev crater, as expected⁹. There is some evidence that both landers were in an updraft during the last few kilometres, but this is not surprising given that the modelling predicted $\sim 40\%$ of the area would be experiencing updrafts at both sites⁹. Perhaps the most basic measure of the atmospheric modelling success is that both landers arrived safely and that the backshell rocket systems^{1,13} on each spacecraft (added partly on account of atmospheric concerns) were both critical to ensuring a safe landing (without them, the Spirit landing in the Gusev crater would have been very close to the limit of the airbag performance envelope).

Thermal inertia

Thermal inertia is a measure of the resistance of surface materials to a change in temperature and can be related to particle size, bulk density and cohesion¹⁴. Surfaces dominated by loose dust have lower thermal inertia and typically high albedo, whereas those dominated by rock or duricrust (cemented soil-like materials) have higher thermal inertia. The fine-component thermal inertia is the thermal inertia of the surface after the thermal radiance attributable to the rocky component is factored out¹⁵.

Orbital thermal inertia measurements of both landing sites^{16–18} suggested surfaces that are competent and load bearing (without thick deposits of fine-grained dust) that pose no special risk to landing or roving¹. The landing location in Gusev crater has a bulk TES thermal inertia of $315 \text{ J m}^{-2} \text{ s}^{-0.5} \text{ K}^{-1}$, which is consistent with the Viking¹⁷- and THEMIS^{5,18}-derived thermal inertias (284 and $306 \text{ J m}^{-2} \text{ s}^{-0.5} \text{ K}^{-1}$, respectively). These thermal inertias suggested the surfaces are dominated by duricrust or cohesionless sand or granules^{19,20}, which is consistent with observed soil characteristics²¹ and Mini-TES measured thermal inertias ($150\text{--}430 \text{ J m}^{-2} \text{ s}^{-0.5} \text{ K}^{-1}$) from the surface²². Average THEMIS thermal inertia along the traverse at the Gusev crater (Fig. 3) varies from $285 \text{ J m}^{-2} \text{ s}^{-0.5} \text{ K}^{-1}$ at the landing site, to $290 \text{ J m}^{-2} \text{ s}^{-0.5} \text{ K}^{-1}$ part of the way up the

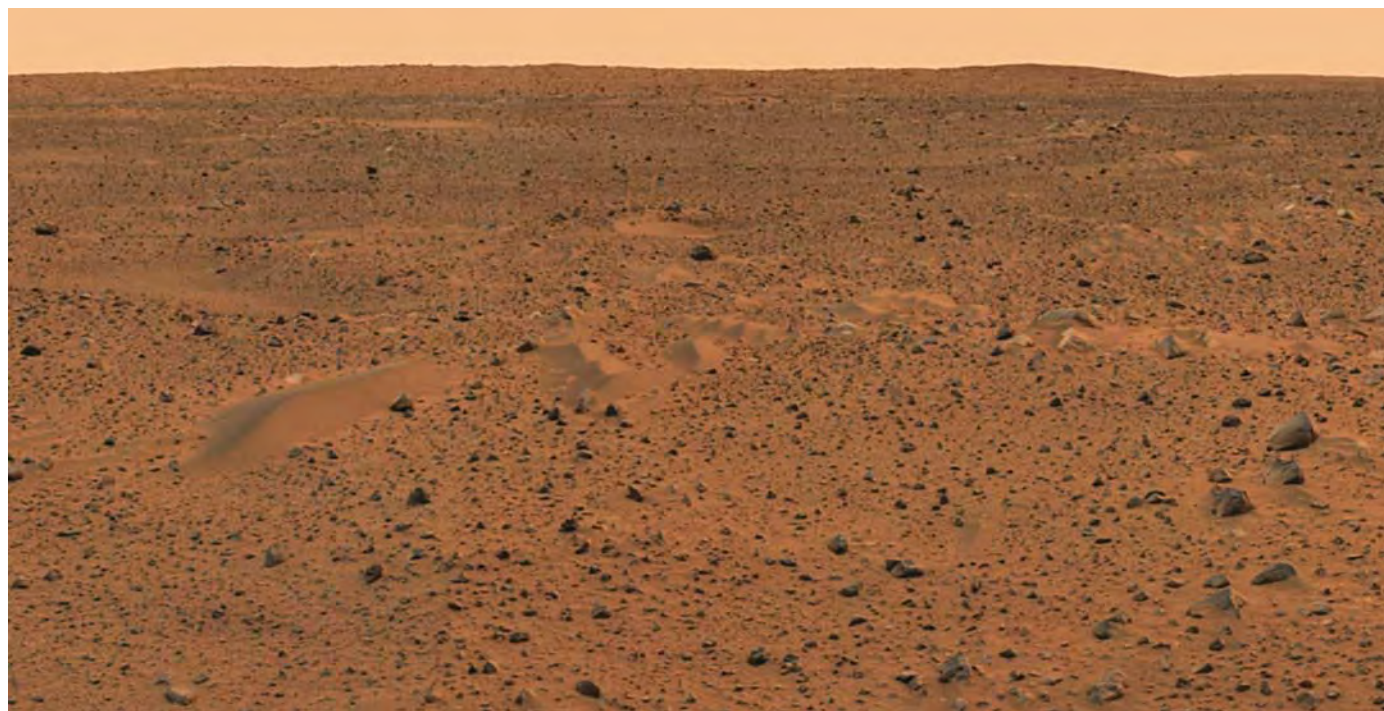


Figure 1 | Portion of the panorama obtained from the Spirit landing site, showing the moderately rocky, relatively smooth plain predicted from remotely sensed data. The bright region on the horizon is the brighter and dustier Bonneville crater rim, which is characteristic of most of the landing ellipse, as opposed to the lower-albedo and less-dusty landing location in a dust devil track. We note filled-in impact craters (circular hollows) and dark

drifts and the pebble-rich surface, consistent with a dark armoured lag or pavement that has relatively little dust. Rock counts from the lander are from this area. This is an approximately true-colour rendering generated from a composite of images acquired through Pancam's 750-nm, 530-nm and 480-nm filters as part of imaging sequences P2215 and P2216 acquired on Spirit sols 4 and 5 (7 and 8 January 2004).

Bonneville ejecta, to $330 \text{ J m}^{-2} \text{ s}^{-0.5} \text{ K}^{-1}$ around Bonneville, and show systematic variations that can be related to observed variations in rock abundance and material properties^{22,23}.

In contrast, the landing location in Meridiani has TES and THEMIS bulk inertias of 200 and $190 \text{ J m}^{-2} \text{ s}^{-0.5} \text{ K}^{-1}$, respectively, although Viking inertias are slightly higher ($\sim 315 \text{ J m}^{-2} \text{ s}^{-0.5} \text{ K}^{-1}$). The TES and THEMIS inertias are similar to the Mini-TES measured inertias of $225 \text{ J m}^{-2} \text{ s}^{-0.5} \text{ K}^{-1}$ and correspond to surfaces dominated by 0.2-mm sand particles²⁰, which is consistent with the ubiquitous fine sand observed at Meridiani²⁴ (Fig. 2).

Albedo and dustiness

Spirit landed in the lowest-albedo portion of the Gusev landing ellipse characterized by dark dust-devil tracks (Fig. 3). As a result, the surface observed at the landing site is substantially less dusty than inferred for the rest of the ellipse. The average TES albedo¹⁶ of the Gusev ellipse is ~ 0.23 and bright areas have albedos as high as 0.26. The low-albedo portion of the ellipse in the dust-devil track region in which Spirit landed has a much lower TES albedo of ~ 0.19 , comparable to the Pancam surface measurement²⁵ (0.20), which is lower than the VL and MPF landing sites. The surface observed by Spirit at the landing site is characterized by a reddish soil surface with many dark granules, pebbles and small rocks as a lag or pavement (Fig. 1) and only modest amounts of bright atmospheric dust coating the rocks and soil surfaces, consistent with the lower albedo and the low dust index for this portion of the ellipse²⁶. The albedos of bright areas like the rim of Bonneville crater that Spirit traversed into are much higher (0.30), consistent with orbital measurements of non-dust-devil track areas.

The average albedo of the Meridiani landing site in orbital data¹⁶ is ~ 0.15 and thus it represents the first landing in a characteristically low-albedo portion of Mars²⁷. Opportunity landed in an area of the ellipse with even lower albedo (~ 0.12) and the dust index of this part of the ellipse is among the lowest on Mars²⁶. The dark sand-rich and dust-free surface observed on the Meridiani plains is consistent with its low albedo (Fig. 2). The brighter rim of the Eagle crater observed in the orbital and descent images is consistent with bright outcrops and brighter red soil surfaces that Opportunity has observed near the Eagle crater rim (Fig. 2). Pancam surface measurements²⁷ yield comparable albedos of 0.12 on the dark plains and higher albedos for the outcrops (0.25) and brighter wind streaks (0.19 to 0.29). The consistency between orbital and surface albedos and the presence or absence of bright dust further supports the use of albedo as a proxy for the dustiness of surfaces on Mars.

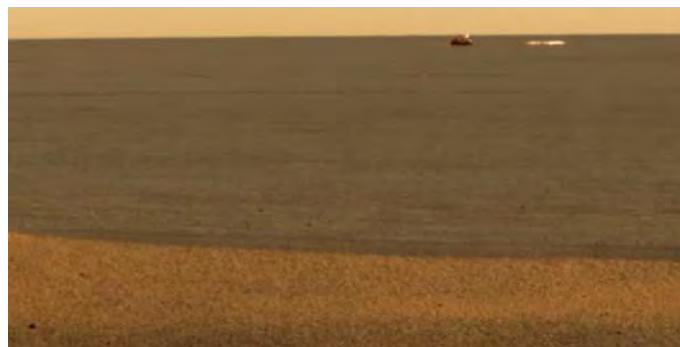


Figure 2 | Image of the Meridiani plain showing its dark, relatively dust- and rock-free plain, as predicted by orbital remote sensing data. The backshell, which is about 1 m high, and the parachute are about 450 m from the rover and illustrate the exceptionally smooth, flat and rock-free plain (except for the bright crater rim in the foreground), which was as predicted before landing. This is an approximately true-colour rendering generated from a composite of images acquired through Pancam's 750-nm, 530-nm and 430-nm filters as part of imaging sequence P2379 acquired on Opportunity sol 21 (14 February 2004).

Rock abundance

The average rock abundance of the Meridiani ellipse is $\sim 5\%$ as estimated from thermal differencing of the Viking Infrared Thermal Mapper (IRTM) data¹⁵. Rock abundance at the Gusev ellipse is higher ($\sim 7\%$) and similar to the global mode of $\sim 8\%$. Opportunity landed at a location near the border of 1% and 6% rock-abundance pixels (1° latitude and longitude) indicating¹⁵ a rock abundance of a few per cent. Spirit is in an 8%-rock-abundance pixel and is not in portions of the ellipse where dense boulder fields were identified in MOC images²⁸. These estimates suggested moderate rock abundance at the Gusev crater and very few rocks at Meridiani Planum, both of which have been relatively benign for driving the rover, as expected.

Rocks greater than ~ 0.04 m in diameter were counted within three roughly 70° sections of panoramas within 10 m of Spirit at the landing site (Mission Success), part of the way up the ejecta (Legacy), and at the rim of Bonneville crater (Bonneville), which have increasing bulk thermal inertias (Fig. 3). Results show that 7%, 5% and 29% of the surface is covered by rocks greater than ~ 0.04 m in diameter (Fig. 4) at these three sites, respectively. The size-frequency distribution of larger rocks (>0.1 m in diameter) generally follows the exponential model distribution based on the VL and MPF landing sites²⁸ for total rock abundances of 5%, 7% and 35% at the three respective sites, although there are far more pebbles at the Spirit landing site (consistent with less bright dust and drift material at this site) than at other locations. The largest rock size increases as the rock abundance increases, from 0.5 m to 0.8 m to 1.3 m in diameter towards the rim of Bonneville crater. Adjusting the intermediate rock count upward to account for the difference in bulk thermal inertia for this location versus the average (290 versus $306 \text{ J m}^{-2} \text{ s}^{-0.5} \text{ K}^{-1}$) (assuming that the difference is due to more rocks)^{23,28}, about 7% of the surface would be covered by rocks more than 0.1 m in diameter, which compares favourably with the IRTM rock abundance¹⁵ estimate of 8%. For effective thermal inertias of rock populations²⁸, the increase in bulk inertia on the Bonneville ejecta blanket is more than explained by the increase in rock

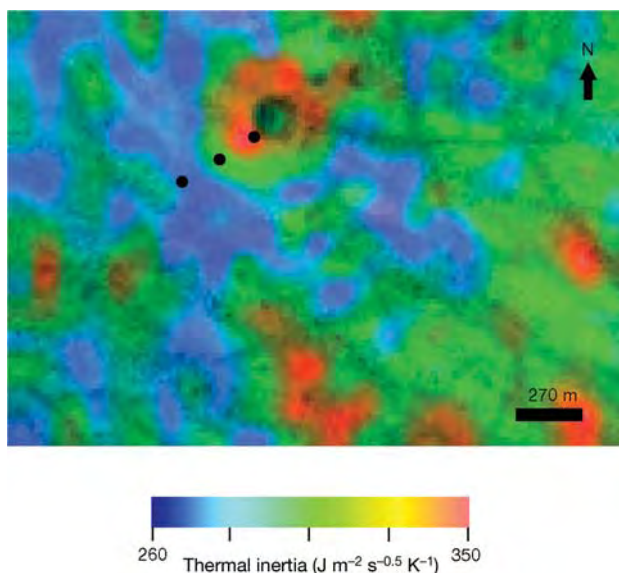


Figure 3 | THEMIS thermal inertia image in colour overlaid on a THEMIS visible image of Spirit landing area. It shows low-albedo, low-thermal-inertia intercrater plains where Spirit landed and locations with higher inertias on the drive to the rim of Bonneville crater where two other full Pancam panoramas²⁵ were acquired. Thermal inertia increases from 270 to $345 \text{ J m}^{-2} \text{ s}^{-0.5} \text{ K}^{-1}$ over this traverse with rock size-frequency distributions (reported in Fig. 4) at these three locations. The southwesternmost black dot is the Mission Success panorama where Spirit landed, the middle black dot is the Legacy panorama part of the way through the ejecta blanket, and the northeasternmost black dot is the rim of the Bonneville crater.

abundance, and suggests a corresponding decrease in the fine-component inertia, which appears consistent with observations of more dust closer to the rim.

The Meridiani plain is effectively devoid of rocks (Fig. 2) and Opportunity is the first lander to sample an area of Mars with very low rock abundance¹⁵. The orbital rock abundance estimate at this site is probably due to the outcrop, which appears to cover roughly 5% of the area within the ~20-m-diameter Eagle crater, and is exposed in crater interiors and rims and in fractures across the plain. In general, the area covered by outcrops and the rock-free plain appears consistent with the orbital estimate of several per cent of the surface covered by rocks at Meridiani Planum.

Slopes

Slopes were evaluated at three length scales important for landing¹: 1 km, 100 m (from MOLA topography) and ≤ 10 m (from Mars Orbiter Camera stereogrammetry and photogrammetry). At all three scales Meridiani Planum is extraordinarily smooth and flat. From Opportunity's traverse telemetry⁸ the root-mean-square (r.m.s.) slopes at these three scales are 0.3°, 0.7° and 1.4°, respectively, and follow a self-affine behaviour with a Hurst exponent²⁹ of 0.64. These slopes are consistent with the slopes reported before landing^{1,30,31} and the exceptionally smooth and flat plain traversed by Opportunity (Fig. 2). The Gusev crater surface appeared rougher than the Meridiani plain, but smoother than VL1 and MPF in orbital data^{1,30,31}, which is consistent with the derived r.m.s. slopes from Spirit of 0.5°, 1.4°, and 2.5° at these three length scales (Hurst exponent of 0.58) and the relatively low-relief plain traversed by Spirit (Fig. 1).

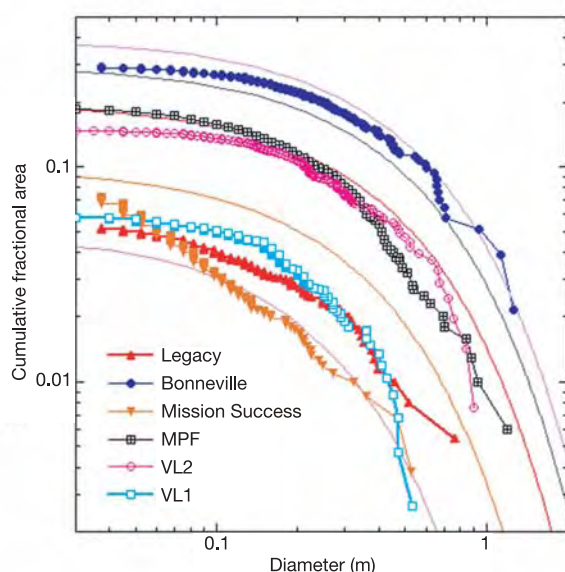


Figure 4 | Rock size-frequency distributions at three locations along the Spirit traverse, VL1, VL2 and MPF landing sites. Locations along the Spirit traverse are shown in Fig. 3. Solid lines are exponential cumulative fractional area versus diameter models²⁸ for 5%, 10%, 20%, 30% and 40% rock abundance (denoted roughly by where curves intersect the ordinate) derived from VL rock counts. Rocks greater than ~0.04 m in diameter were counted in roughly 70° sectors of panoramas within 10 m of the lander at Mission Success (1,089 rocks in 56.9 m² area from 0° to 76°), Legacy (426 rocks in 58 m² area from 318° to 28°) and Bonneville (689 rocks in 84.1 m² area from 255° to 353°). Rocks in the panoramas were manually identified using an interactive Graphic User Interface of RockIT, a component of the OASIS software (Onboard Autonomous Science Investigation System)³⁹. Range data were then used to calculate apparent width (1.33 times the diameter²⁸) and height. A rock of apparent width 5 cm is easily resolved at 10 m by Pancam (~18 pixels wide)²⁵, so that errors associated with these measurements²⁸ are not important on the log-log plot.

Radar

Radar reflectivity values of 0.05 and 0.04 evaluated before landing¹ indicated surfaces with loosely constrained, but reasonable, bulk densities of ~1,500 and ~1,200 kg m⁻³ at Meridiani and Gusev, respectively, that pose no special problem to landing or roving¹⁹ and are similar to the range of bulk densities of soils that were successfully landed on and roved over by Mars Pathfinder³². Preliminary processing of later near-nadir 3.5-cm backscatter data with much higher spatial resolution (5 km × 5 km versus 10 km × 150 km) yield somewhat lower reflectivities of 0.02 ± 0.01 at both landing sites³³, which might be due to Doppler- and range-aliasing into the near-nadir quasi-specular echo that produces an elevated apparent noise level and reduced reflectivity. In any case, load-bearing surfaces have been confirmed by the successful landing and roving at the two sites.

The r.m.s. slope or roughness derived using the Hagfors model^{34,35} indicated a smoother surface at Meridiani than at MPF (3.5-cm r.m.s. 1.4° versus 4.5°) and a smoother surface at Gusev than at VL1 (12.6-cm r.m.s. 1.7° versus 6°)¹. Interpretation of radar data predicted that Meridiani Planum would be much less rocky and smoother than the VL2 site, and that the Gusev crater would have a combination of roughness at decimetre scales similar to or greater than VL1 and MPF sites, but would be smoother at metre scales¹. These predictions appear consistent with the very flat, rock-free plain at Meridiani and the generally smooth, moderately rocky surface at the Gusev crater, where r.m.s. slopes from Front Hazcam stereo pairs average 3° at a 3-m scale for both rovers, but average about 30° for Spirit and 20° for Opportunity at a 10-cm scale.

Results

The close correspondence between surface characteristics inferred from orbital remote sensing data and that found at the landing sites argues that future efforts to select safe landing sites will be successful. Linking the five landing sites to their remote sensing signatures suggests that they span many of the important, probably safe surfaces available for landing on Mars, which have moderate to high thermal inertia with low to high albedo (but not low albedo and low thermal inertia). Our results show that basic engineering parameters important for safely landing spacecraft such as elevation, atmospheric profile, bulk density, rock distribution and slope can be adequately constrained using available and targeted remote sensing data.

In contrast to accurately defining the important physical characteristics of the surface, geological interpretations of the sites were less successful with respect to addressing the main scientific objectives of the mission (preserving evidence of an aqueous environment). The TES haematite signature and the geological setting of Meridiani inferred from THEMIS did correctly predict the origin of the haematite as a low-temperature precipitate⁶ and the discovery of sulphate evaporites formed in an ancient aqueous environment³⁶. However, the cratered plains inside Gusev do not appear to be sedimentary rocks deposited in a crater lake fed by Ma'adim Vallis, but instead appear to be a volcanic (basalt) surface that has been dominated by impact and eolian activity^{37,38}. This demonstrates the uncertainty in predicting precisely what geologic materials will be available for study at landing sites from remotely sensed data.

1. Golombek, M. P. *et al.* Selection of the Mars Exploration Rover landing sites. *J. Geophys. Res.* **108**(E12), 8072, doi:10.1029/2003JE002074 (2003).
2. Cabrol, N. A., Grin, E. A. & Landheim, R. Ma'adim Vallis evolution: geometry and models of discharge rate. *Icarus* **132**, 362–377 (1998).
3. Kuzmin, R., *et al.* Geologic map of the MTM-15182 and MTM-15187 quadrangles, Gusev crater-Ma'adim Vallis region, Mars. *US Geol. Surv. Map I-2666* (US Geological Survey, Washington DC, 2000).
4. Arvidson, R. E. *et al.* Mantled and exhumed terrains in Terra Meridiani, Mars. *J. Geophys. Res.* **108**(E12), 8073, doi:10.1029/2002JE001982 (2003).
5. Christensen, P. R. *et al.* Morphology and composition of the surface of Mars: Mars Odyssey THEMIS results. *Science* **300**(5628), 2056–2061 (2003).
6. Christensen, P. R. & Ruff, S. W. Formation of the hematite-bearing unit in Meridiani Planum: evidence for deposition in standing water. *J. Geophys. Res.* **109**, E08003, doi:10.1029/2003JE002233 (2004).

7. Arvidson, R. E. *et al.* Localization and physical properties experiments conducted by Spirit at Gusev crater. *Science* **305**, 821–824 (2004).
8. Arvidson, R. E. *et al.* Localization and physical properties experiments conducted by Opportunity at Meridiani Planum. *Science* **306**, 1730–1733 (2004).
9. Kass, D. M. *et al.* Analysis of atmospheric mesoscale models for entry, descent, and landing. *J. Geophys. Res.* **108**(E12), 8090, doi:10.1029/2003JE002065 (2003).
10. Magalhaes, J. A., Schofield, J. T. & Seiff, A. Results of the Mars Pathfinder atmospheric structure investigation. *J. Geophys. Res.* **104**, 8943–8956 (1999).
11. Smith, M. D. Interannual variability in TES atmospheric observations of Mars during 1999–2003. *Icarus* **167**, 148–165 (2004).
12. Smith, D. E. *et al.* Mars Orbiter Laser Altimeter (MOLA): Experiment summary after the first year of global mapping of Mars. *J. Geophys. Res.* **106**, 23689–23722 (2001).
13. Crisp, J. A. *et al.* Mars Exploration Rover mission. *J. Geophys. Res.* **108**(E12), 8061, doi:10.1029/2002JE002038 (2003).
14. Christensen, P. R. & Moore, H. J. in *Mars* (eds Kieffer, H. H. *et al.*) 686–727 (Univ. Ariz. Press, Tucson, 1992).
15. Christensen, P. R. The spatial distribution of rocks on Mars. *Icarus* **68**, 217–238 (1986).
16. Mellon, M. T., Jakosky, B. M., Kieffer, H. H. & Christensen, P. R. High-resolution thermal inertia mapping from the Mars Global Surveyor Thermal Emission Spectrometer. *Icarus* **148**, 437–455 (2000).
17. Kieffer, H. H. *et al.* Thermal and albedo mapping of Mars during the Viking Primary Mission. *J. Geophys. Res.* **82**, 4249–4291 (1977).
18. Fergason, R. L. & Christensen, P. R. Thermal inertia using THEMIS infrared data. *Lunar Planet. Sci. XXIV* abstr. 1785 (CD-ROM, Lunar and Planetary Institute, Houston, 2003).
19. Golombek, M. P. *et al.* Selection of the Mars Pathfinder landing site. *J. Geophys. Res.* **102**, 3967–3988 (1997).
20. Presley, M. A. & Christensen, P. R. Thermal conductivity measurements of particulate materials. 2. Results. *J. Geophys. Res.* **102**, 6551–6566 (1997).
21. Herkenhoff, K. E. *et al.* Textures of the soils and rocks at Gusev crater from Spirit's Microscopic Imager. *Science* **305**(5685), 824–826 (2004).
22. Christensen, P. R. *et al.* Initial results from the Mini-TES experiment in Gusev crater from the Spirit rover. *Science* **305**(5685), 837–842 (2004).
23. Moersch, J. E., *et al.* Comparison of orbital infrared observations and surface measurements by the Mars Exploration Rover Spirit at Gusev crater. *Lunar Planet. Sci. XXXVI* abstr. 2020 (CD-ROM, Lunar and Planetary Institute, Houston, 2005).
24. Herkenhoff, K. E. *et al.* Evidence for ancient water on Meridiani Planum from Opportunity's Microscopic Imager. *Science* **306**, 1727–1730 (2004).
25. Bell, J. F. III *et al.* Pancam multispectral imaging results from the Spirit rover at Gusev crater. *Science* **305**(5685), 800–806 (2004).
26. Ruff, S. W. & Christensen, P. R. Bright and dark regions on Mars: Particle size and mineralogical characteristics based on Thermal Emission Spectrometer data. *J. Geophys. Res.* **107**(E12), 5127 doi:10.1029/2001JE001580 (2002).
27. Bell, J. F. *et al.* Pancam multispectral imaging results from the Opportunity rover at Meridiani Planum. *Science* **306**, 1703–1709 (2004).
28. Golombek, M. *et al.* Rock size-frequency distributions on Mars and implications for MER landing safety and operations. *J. Geophys. Res.* **108**(E12), 8086, doi:10.1029/2002JE002035 (2003).
29. Shepard, M. K. *et al.* The roughness of natural terrain: A planetary and remote sensing perspective. *J. Geophys. Res.* **106**, 32777–32795 (2001).
30. Anderson, F. S. *et al.* Analysis of MOLA data for the Mars Exploration Rover landing sites. *J. Geophys. Res.* **108**(E12), 8084, doi:10.1029/2003JE002125 (2003).
31. Kirk, R. *et al.* High-resolution topomapping of candidate MER landing sites with Mars Orbiter Camera narrow angle images. *J. Geophys. Res.* **108**(E12), 8088, doi:10.1029/2003JE002131 (2003).
32. Moore, H. J. *et al.* Soil-like deposits observed by Sojourner, the Pathfinder rover. *J. Geophys. Res.* **104**, 8729–8746 (1999).
33. Larsen, K. W., Haldemann, A. F. C., Jurgens, R. F. & Slade, M. A. Radar observations of recent Mars landing sites. *Lunar Planet. Sci. XXXV* abstr. 1050 (Lunar and Planetary Institute, Houston, 2004).
34. Hagfors, T. Backscattering from an undulating surface with applications to radar returns from the Moon. *J. Geophys. Res.* **69**, 3779–3784 (1964).
35. Evans, J. V. & Hagfors, T. *Radar Astronomy* 620 (McGraw-Hill, New York, 1968).
36. Squyres, S. W. *et al.* In-situ evidence for an ancient aqueous environment on Mars. *Science* **306**, 1709–1714 (2004).
37. Grant, J. A. *et al.* Surficial deposits at Gusev crater along Spirit rover traverses. *Science* **305**, 807–810 (2004).
38. Squyres, S. W. *et al.* The Spirit rover's Athena science investigation at Gusev crater, Mars. *Science* **305**, 794–799 (2004).
39. Castaño, R., *et al.* Autonomous onboard traverse science system. *Proc. 2004 IEEE Aerospace Conf. (Big Sky, Montana, March 2004)* abstr. 1375 (Institute of Electrical and Electronics Engineers, New York, NY, 2004).

Acknowledgements We are deeply indebted to the MER engineers for offering us the opportunity to test our landing site predictions with data from the surface of Mars. We acknowledge the contributions of R. Blanchard, P. Withers and the MER Atmospheric Advisory Team to the interpretations of the atmospheric entry information. M. Wyatt provided the THEMIS thermal inertia image. R. Castano, A. Castano, B. Bornstein and R. C. Anderson developed OASIS, R. Deen provided panoramas and associated range at correct resolution, and T. Stough and M. Judd provided rock counts. Research described in this paper was done by the MER project, Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.P.G. (mgolombek@jpl.nasa.gov).

An integrated view of the chemistry and mineralogy of martian soils

Albert S. Yen¹, Ralf Gellert², Christian Schröder³, Richard V. Morris⁴, James F. Bell III⁵, Amy T. Knudson⁶, Benton C. Clark⁷, Douglas W. Ming⁴, Joy A. Crisp¹, Raymond E. Arvidson⁸, Diana Blaney¹, Johannes Brückner², Philip R. Christensen⁶, David J. DesMarais⁹, Paulo A. de Souza Jr¹⁰, Thanasis E. Economou¹¹, Amitabha Ghosh¹², Brian C. Hahn¹³, Kenneth E. Herkenhoff¹⁴, Larry A. Haskin⁸, Joel A. Hurowitz¹³, Bradley L. Joliff⁸, Jeffrey R. Johnson¹⁴, Göstar Klingelhöfer³, Morten Bo Madsen¹⁵, Scott M. McLennan¹³, Harry Y. McSween¹², Lutz Richter¹⁶, Rudi Rieder², Daniel Rodionov³, Larry Soderblom¹⁴, Steven W. Squyres⁵, Nicholas J. Tosca¹³, Alian Wang⁸, Michael Wyatt⁶ & Jutta Zipfel²

The mineralogical and elemental compositions of the martian soil are indicators of chemical and physical weathering processes. Using data from the Mars Exploration Rovers, we show that bright dust deposits on opposite sides of the planet are part of a global unit and not dominated by the composition of local rocks. Dark soil deposits at both sites have similar basaltic mineralogies, and could reflect either a global component or the general similarity in the compositions of the rocks from which they were derived. Increased levels of bromine are consistent with mobilization of soluble salts by thin films of liquid water, but the presence of olivine in analysed soil samples indicates that the extent of aqueous alteration of soils has been limited. Nickel abundances are enhanced at the immediate surface and indicate that the upper few millimetres of soil could contain up to one per cent meteoritic material.

The 1976 Viking landers^{1,2} provided the first elemental analyses³ of martian surface materials. These results, using X-ray fluorescence spectrometers, indicate a mafic composition of the soils and a level of sulphur two orders of magnitude higher than the average crust of Earth⁴. Two decades after the Viking missions, Mars Pathfinder⁵ arrived with an Alpha Proton X-ray Spectrometer (APXS)⁶ capable of identifying elements below the detection limit of the Viking X-ray fluorescence spectrometer. More importantly, the Pathfinder APXS was mounted on a mobile platform, enabling analyses of rock surfaces as well as soils. Compositional averages showed that soils were significantly enhanced in Fe and Mg relative to the rocks, suggesting that soil compositions are not dominated by the physical weathering products of local rocks^{7–9}. The Viking and Pathfinder landers were also equipped with multispectral imagers^{10,11}, which confirmed orbital and Earth-based observations of ferric iron absorptions, indicative of oxidized surface materials.

In January 2004, the Mars Exploration Rovers (MERs) Spirit and Opportunity landed in Gusev crater¹² and on the haematite-rich plains of Meridiani Planum¹³, respectively. The science payload of each rover consists of the Panoramic camera (Pancam)¹⁴, the Miniature Thermal Emission Spectrometer (Mini-TES)¹⁵, the Microscopic Imager¹⁶, the Mössbauer spectrometer¹⁷, the APXS¹⁸, the Rock Abrasion Tool¹⁹, and a suite of magnets²⁰. The use of identical sets of complementary instruments at the two landing sites enables a thorough investigation of martian soils.

Primarily on the basis of morphology evident in Microscopic Imager images, the soils at Gusev crater can be categorized into four

components (Table 1 and Fig. 1). A thin, ~1-mm-thick layer of easily compacted, fine-grained, 'bright dust' is found at the immediate surface. Beneath this layer is a 'dark soil' with a grain size of up to 100 µm, just at the limit of the Microscopic Imager resolution. Imprints produced by the Mössbauer spectrometer upon contact with these dark soils indicate that they also contain a significant population of smaller grains²¹. Aeolian (wind-deposited) 'bedform armour' consists of millimetre-sized grains, and larger 'lithic fragments' are embedded in the soil.

Four components of the soil are also present at Meridiani Planum (Table 1 and Fig. 1). Haematite-rich 'spherules' and their fragments are present throughout this landing site²² in amounts that can be detected from orbit²³. 'Clasts' of variable angularity and vesiculation are interspersed among the spherules²⁴. Excluding spherules and other clasts, the fine-grained deposits at the surface are dominated by a 'dark soil' with a maximum grain size of approximately 100 µm. 'Bright dust' is present in small patches at the surface as well as in subsurface deposits exposed by the rover wheels.

Bright dust

The elemental composition of the bright surface dust at Gusev crater is remarkably uniform²⁵. Undisturbed soils that do not include pebbles or other rock fragments (Table 1) have variations in major and minor elements that are less than 15% of the average value. The combination of Mössbauer and APXS data establishes that the nanophase iron oxide²⁶ component of soils is closely associated with the occurrence of sulphur (Fig. 2a). Furthermore, the negative

¹Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA. ²Max Planck Institut für Chemie, ³Johannes Gutenberg University, D-55128 Mainz, Germany. ⁴NASA Johnson Space Center, Houston, Texas 77058, USA. ⁵Cornell University, Department of Astronomy, Ithaca, New York 14853, USA. ⁶Arizona State University, Department of Geological Sciences, Tempe, Arizona 85287, USA. ⁷Lockheed Martin Corporation, Littleton, Colorado 80127, USA. ⁸Washington University, Saint Louis, Missouri 63130, USA. ⁹NASA Ames Research Center, Moffett Field, California 94035, USA. ¹⁰Companhia Vale do Rio Doce, 29030-900 Rio de Janeiro, Brazil. ¹¹Enrico Fermi Institute, University of Chicago, Chicago, Illinois 60637, USA. ¹²Department of Earth and Planetary Sciences, University of Tennessee, Knoxville, Tennessee 37996, USA. ¹³State University of New York, Department of Geosciences, Stony Brook, New York 11794, USA. ¹⁴US Geological Survey, Flagstaff, Arizona 86001, USA. ¹⁵Niels Bohr Institute, University of Copenhagen, DS-2100 Copenhagen, Denmark. ¹⁶DLR Institut für Raumsimulation, D-51170 Cologne, Germany.

correlation between sulphur and olivine indicates that this ferrous iron component decreases as the amount of sulphur increases (Fig. 2b), firmly refuting suggestions of abundant ferrous sulphate in martian soils²⁷. These results are consistent with the expectation that the sulphur component of the soil (a proxy for alteration) is associated with the most weathered and oxidized mineral phases.

At Meridiani, the majority of surface soils consists of clasts, spherules and/or dark soil, but there have been two measurements of bright surface dust (Table 1). Compositions of these targets are within 15% of the average of bright dust at Gusev crater for all major and minor elements, with the exception of Na (22%) and Cr (28%). The variability in Na might result from differences in the low energy threshold of the two APXS instruments, and the Cr variation could be attributable to the small sample set and the low count rates for an oxide present at ~0.3 weight per cent. Subsurface soils at Meridiani exposed by the rover wheels also provide an excellent match to the bright surface dust deposits. Major- and minor-element abundances of Big Dig/Hema Trench 1 (Table 1) are within 15% of the average bright surface dust composition. PhotoTIDD/Nougat is also similar but the relatively volatile elements S and Cl are 20% and 27% lower, respectively, than the average bright surface dust. These subsurface soils probably represent an earlier episode of dust deposition that has since been covered by the influx of the dark soil.

The elemental chemistry of the soils measured at the five landing sites on Mars are plotted in Fig. 3. The bright dust at Gusev crater and the Meridiani plains plot in relatively tight clusters with respect to each other and are distinct from the rock compositions. Systematic offsets in Viking Mg, Al, and Ti could represent actual variability in the soil chemistry or the absolute accuracy of the instruments. Nonetheless, the observed chemistry of the surface soil is remarkably similar given the separation of the landing sites and the differences in the instrument hardware.

Pancam spectra of bright dust deposits (Fig. 4) are similar at Gusev and Meridiani, and both are similar to bright dust spectra measured over the same wavelengths by Mars Pathfinder²⁸ and telescopic observations²⁹. The spectra are consistent with a composition dominated by nanophase ferric oxides³⁰. Differences in the absolute reflectivity of the bright soils at the MER landing sites could result from a smaller mean particle diameter at Meridiani relative to Gusev, to the presence of an additional spectrally neutral component in the Meridiani dust, and/or to differences in surface texture.

Bright, undisturbed soils at Gusev have a Mini-TES spectral signature similar to that of Mars Global Surveyor TES spectra of regions of Mars^{31,32} with high albedo and low thermal inertia. A

bright streak downwind of Eagle crater at the Meridiani site exhibits a spectral signature that also matches Mars Global Surveyor TES global dust (Fig. 5)³³. This remarkable consistency indicates that local dust deposits have the same homogeneous composition as the global average Mars dust.

The bright dust deposits at Gusev and Meridiani have similar physical properties and the dust readily adheres to the contact plate in front of the Mössbauer spectrometer, resulting in extraction of clods (Fig. 1d and e). In addition, the magnetic properties investigation on both rovers indicates that all dust particles are magnetic and that they have a composition consistent with the bright dust^{34,35}. These data, taken collectively, indicate that the thin layer of bright surface dust is a global soil unit with distinct compositional and physical properties.

Dark soil

The dark soils are low-sulphur endmembers. With the exception of haematite-rich soils at Meridiani, which have increased levels of iron³⁶ and the interiors of trenches, other soils at Gusev and Meridiani plot on a line between the dark soil and bright dust (Fig. 3). The dark soils at the two landing sites are reasonably consistent with each other, and ratios of dark soil (Table 1) to bright surface dust exhibit similar profiles (Fig. 6). The large discrepancy in Br results from the location of the soil units at the two MER sites. At Meridiani, the plotted samples of dark soil and bright dust are found at the immediate surface, and Br is at the instrument detection limit, so the ratio is that of small numbers. At Gusev, the dark soil is found beneath the immediate surface, where Br is enriched (see discussion below).

A direct comparison of Mössbauer spectra shows that the dark soil targets at the two sites are essentially identical in iron mineralogy and dominated by olivine and pyroxene (Supplementary Fig. A). The mean percentages of iron in olivine, pyroxene, nanophase iron oxide, and magnetite are 38%, 38%, 15% and 9%, respectively²², for the dark soil targets listed in Table 1. The standard deviations of the four iron minerals across these five samples are 1.6%, 2.8%, 1.9% and 4.3%. The variability in magnetite results from the presence of grains of magnetite-rich bedform armour (Table 1) in the Bear Paw/Panda New target. All other variations are near the $\pm 2\%$ absolute accuracy of the fits to the Mössbauer data and are small relative to the overall variability in the soils^{22,26}.

Pancam spectra of dark soil deposits are also similar at Gusev and Meridiani (Fig. 4). The dark soil spectra at both MER sites are similar to Pathfinder and telescopic data of dark soils and low-albedo regions^{28,29} in that they exhibit a weak ferric absorption edge

Table 1 | Endmember components of martian soils

Site	Soil component	Description	Figure	Representative APXS/Mössbauer Targets*	Sol(s)
Gusev crater	Bright dust	Global unit	1a	Gusev/First Soil Sugar Loaf Flats/Soil 1 Deserts/Gobi 1 Truckin Flats/Accelerator	14 65 68–71 126
	Dark soil	Similar to dark soil at Meridiani	1b	Bear Paw/Panda New Santa Anita/Seattle Slew Shredded/Dark 4	73–74 135 158
	Bedform armour	Abundant magnetite	1c	Arena/Crest Angel Flats/Halo 01	41 45
	Lithic fragments	Abundant magnetite	1d	Ramp Flats/Soil 1 Wrinkle/Ridge 1 (Mössbauer only)	44 54
	Bright dust	Global unit	1e	Mont Blanc/Les Hauches (surface) Hilltop/McDonnell (surface) Big Dig/Hema Trench 1 (subsurface) PhotoTIDD/Nougat (subsurface)	59–60 123 24–25 89–90
Meridiani Planum	Dark soil	Similar to dark soil at Gusev	1f	Millstone/Dahlia Auk/Auk RAT	165–167 237–238
	Spherules	Haematite concretions	1g	Dog Park/Jack Russell PhotoTIDD/Fred Ripple	80 91
	Clasts (mostly angular/vesiculated)	Possibly basaltic	1h	Not yet analysed	–

* MER APXS data used in analyses are tabulated in Supplementary Tables A, B and C.

indicative of the presence of altered iron-bearing minerals. However, in the near-infrared the MER dark soil data are different from average Pathfinder or telescopic data. Specifically, MER dark soils exhibit a shallow and broad absorption band centred near 900 nm that is probably due to the presence of ferrous-iron bearing silicates (for example, pyroxene)³⁷. A similar band at the same position is observed in Pancam spectra of dust-poor rock surfaces at Gusev³⁸. Thus, the dark soils at both MER sites appear to contain a significant component of less-altered mafic material, consistent with the Mössbauer results.

The variability observed in the Mini-TES spectra of dark soils at Gusev is dominated by contributions of the ubiquitous dust. Linear deconvolution of dark soil spectra from a rover track, normalized to remove the dust component, indicates a suite of basaltic minerals: ~45% pyroxene, ~35% plagioclase feldspar, ~15% olivine, 5% glass, and less than 5% sulphates or oxides. These compositions are the same as previously reported results³² with the exception of glass, which was not included as an endmember in the earlier analyses. Dark soils at Meridiani are spectrally similar to those at Gusev (Fig. 5). A representative target called Auk in Endurance crater has a basaltic composition consisting of ~35% pyroxene, ~40% plagioclase feldspar, ~10% olivine, ~15% glass, and less than 5%

sulphates and oxides, a result similar to previously analysed haematite-poor dark soil in Eagle crater³³. The accuracy of Mini-TES mineral retrievals are ± 5 –10% (ref. 32), which is of the order of the variation in mineral abundances between the two landing sites. Dark soils at both MER sites are well matched by Mars Global Surveyor TES data of low-albedo, globally common, basaltic surfaces on Mars^{39–41}.

Other soil components

The surfaces of aeolian bedforms on the Gusev plains are armoured with rounded, well-sorted, millimetre-sized grains (Fig. 1c). Mössbauer spectra of these targets are significantly enhanced in magnetite, by over a factor of two in certain cases, relative to bright dust and dark soil. Subrounded pebbles (Fig. 1d) also exhibit magnetite enrichments and probably have a similar origin. The elemental composition of the bedform armour indicates that the grains are sorted fragments of Gusev plains basalts. Relative to the average composition of the surface dust, these grains are enhanced in Fe, Ca and Cr and depleted in Ti, Ni, Zn, S, Cl and K, as is expected for a mixture of Gusev basalts with global dust. That is, the abundance of these elements in the bedform armour is between that of the dust and the Gusev basalts⁴². Mg and Br, however, do not conform to this interpretation. The Mg abundance in the bedform armour is comparable to that in the dust, but the amount of Mg in rocks is ~20% higher. This apparent loss of Mg in the bedform armour could be explained by the preferential retention of heavy minerals such as magnetite relative to the olivine phases in the original rocks⁴³. This possibility is consistent with the increase in the magnetite to olivine ratio in the Mössbauer measurements of bedform armour as compared to the same ratio in plains basalts. Concentrations of Br in the bedform armour can be more than twice as high as in typical surface soils. An explanation for this enrichment is discussed below.

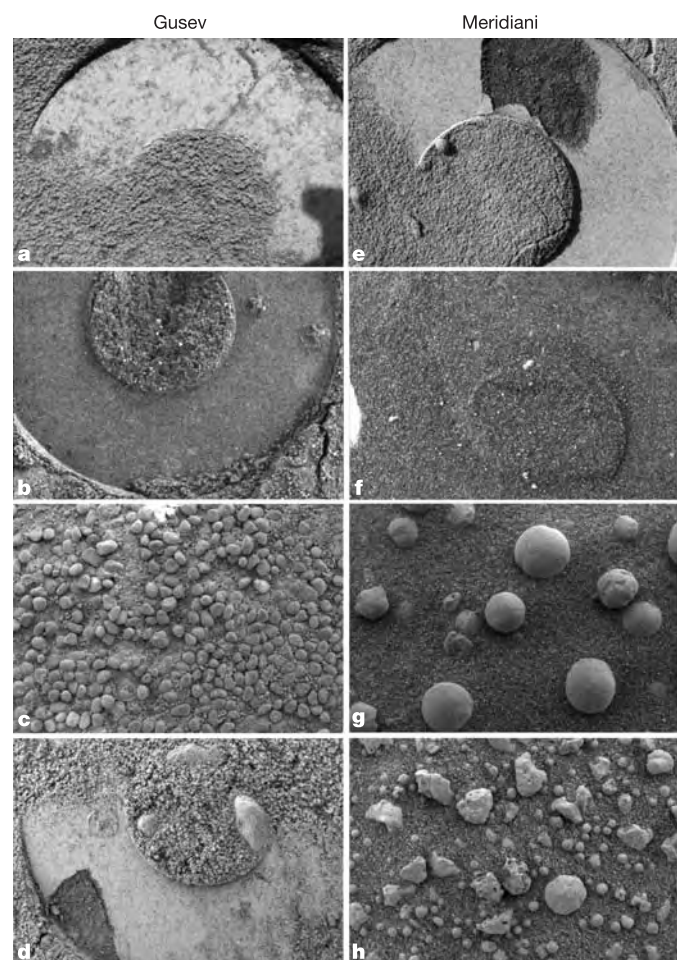


Figure 1 | Microscopic Imager images, each 3 cm across. a–d, Gusev crater images: a, Bright dust (sol 65); b, dark soil (sol 158); c, millimetre-sized bedform armour (sol 41); d, rounded pebbles in a matrix of surface dust (sol 54). Meridiani Planum images: e, Bright dust (sol 123); f, dark soil (sol 167); g, haematitic spherules on a bed of dark soil (sol 14); h, sub-angular, vesicular clasts on dark soil (sol 53). All images except g and h show an imprint of the annular Mössbauer contact plate. In d and e, small patches of soil adhered to the Mössbauer contact plate, revealing underlying dark soil.

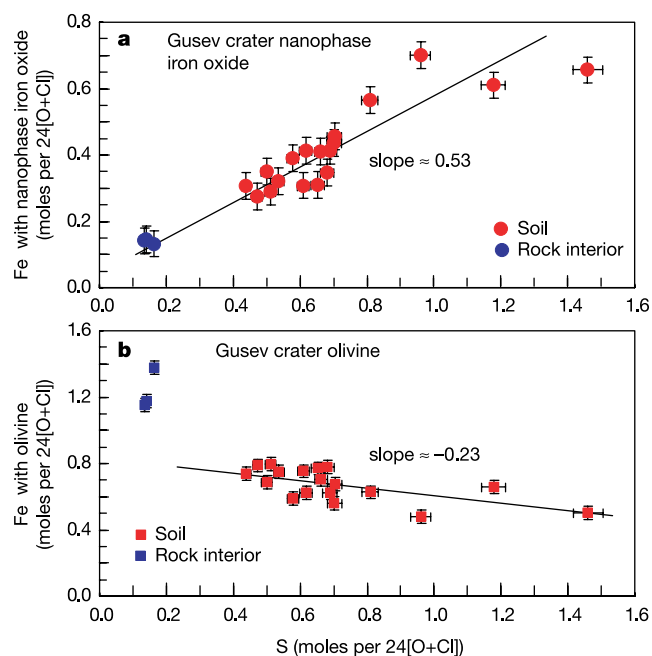


Figure 2 | Correlations between APXS elemental chemistry and iron mineral phases measured by Mössbauer²⁶ in Gusev soils. Molar concentrations with respect to the number of anions are shown. a, Positive correlation indicates that nanophase iron oxide is a carrier of S. The low Fe:S ratio (~1:2) suggests that S is also present in phases which do not contain iron. b, Negative correlation between iron in olivine and the S concentration is consistent with olivine being an unweathered mineral and S associating with altered phases. Error bars represent 2-sigma statistical errors in the APXS data and fitting uncertainties in Mössbauer data.

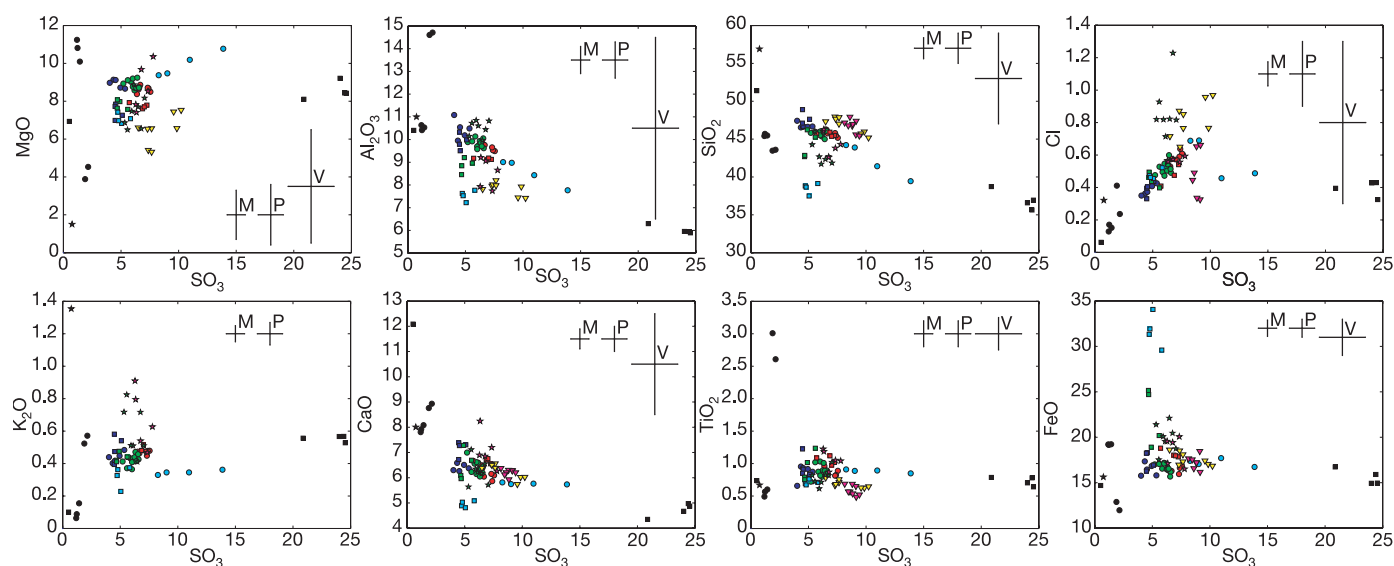


Figure 3 | Composition of martian surface materials. Circles, squares, stars and triangles represent Spirit, Opportunity, Pathfinder^{7,8} and Viking³ data, respectively. MER data: black (rocks), red (bright dust), blue (dark soil), cyan (haematitic soils at Meridiani, high-sulphur trench interiors at Gusev), green (other soils). Pathfinder data in magenta⁷ and green⁸ represent

independent fits of the same data; black is the sulphur-free rock composition⁷. Viking 1 and 2 data are plotted in yellow and magenta, respectively. Renormalization uses iron as FeO and average Gusev values for elements not measured. Approximate error bars representing the uncertainty shown for MER ('M'), Pathfinder ('P'), and Viking ('V') data.

The spherules in rocks and soils at Meridiani Planum are clearly enriched in haematite²². Ratios of APXS data on the spherules relative to a spherule-free background show the expected increase in iron content as well as a decrease in most other elements resulting from dilution (for example, more haematite means less silicates in the field of view). In spherule-rich targets, the abundance of Ni correlates with Fe, indicating that these cations exhibit similar chemical responses during the formation of the spherules.

Origin of the soils

At Gusev crater, the bright dust and dark soil have substantially greater concentrations of S, Cl, K, Ti, Ni and Zn relative to the

analysed rocks⁴² in the plains (Supplementary Fig. B). The increases in S, Cl and Zn could be attributed to precipitates of volcanic outgassing⁴⁴, but variations in other elements are difficult to explain without a significant contribution of material with compositions different from that of the plains basalts.

At Meridiani, the bright dust and dark soil components have sulphur levels comparable to that of bright dust and dark soil at Gusev, yet the local rock outcrops have sulphur concentrations a factor of 4 or 5 larger. From the Mössbauer data, jarosite is not detected in Meridiani soils, and a maximum of only 1% olivine is detected in abraded outcrop rocks²². Therefore, the bright surface dust and the dark soil have been transported to Meridiani Planum by wind-related processes. A similar situation probably applies at Gusev.

There are clearly basaltic fragments at Gusev and haematitic spherules in Meridiani soils that originated from the local rocks. However, the available compositional data indicate that outcrop rocks at Meridiani and plains basalts at Gusev do not contribute significantly to the surrounding bright surface dust and dark soil. This interpretation is further supported by Fig. 3, which shows that soil compositions at five landing sites on Mars are more similar to each other than to the analysed rocks.

The extent of aqueous alteration of the soil at both MER sites has been rather limited. In contrast to Meridiani outcrop rocks that have ferric to total iron ratios in excess of 0.84 (ref. 22), soils that do not include spherules have ferric to total iron ratios of less than 0.42. The one exception is the floor of Big Dig/Hema Trench 1, which has $\text{Fe}^{3+}/\text{Fe}_{\text{total}} = 0.52$. The bottoms of the Gusev and Meridiani trenches, exhibiting the highly oxidized, sulphur-rich soils, still have 19% to 26% of the iron in olivine. This presence of olivine indicates that the soils at depth are either only partially weathered or result from mixing with olivine-rich soils.

Bromine concentrations at Meridiani and Gusev soils are typically less than ~50 p.p.m. at the surface and elevated (factors of 2 to 30 higher) in bedform armour, low-lying rocks, and subsurface soils. Compounds that contain Br are among the most soluble mineral phases, and thus its presence is a probable indicator of liquid water activity. Given the association with high-thermal-inertia materials and subsurface cold traps, the observed behaviour of Br is consistent

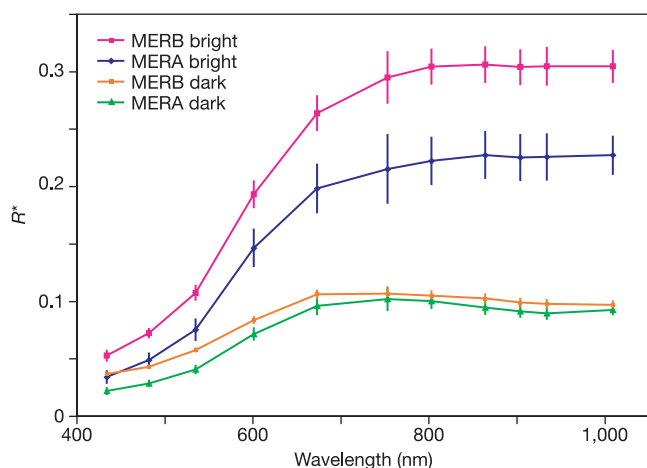


Figure 4 | Average Pancam 11-colour spectra of bright surface dust and dark soil at the Gusev and Meridiani sites. The parameter R^* is the brightness of the surface divided by the brightness of the Pancam radiometric calibration target scaled to its equivalent Lambert reflectance. Error bars represent the variance of all the spectra used to generate the average value plotted.

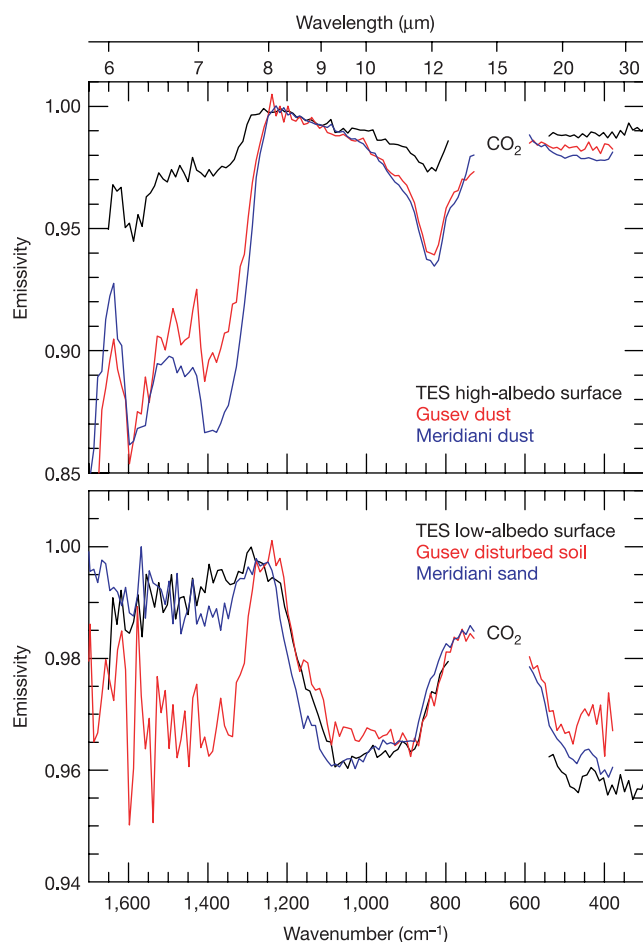


Figure 5 | Comparison of orbital (TES) and surface (Mini-TES) thermal infrared spectra. Bright dust (top): Gusev ('Serpent', sol 70); Meridiani ('CoolWhip', sol 57). Due to averaging over multiple incidence angles, the TES signature has lower spectral contrast³¹. Dark soil (bottom): Gusev ('Skid', sol 89); Meridiani ('Auk', sol 197). A contribution from bright dust³² is present in the Gusev dark soil spectrum, producing differences in the 6–8- μm region and the negative slope between 8 and 12 μm . The broader absorption in the 8–12- μm region of the Meridiani dark soil is attributable to additional sulphate components³³.

with mobilization under climatic conditions similar to present-day Mars or during periods in the obliquity cycle where the mixing ratio of atmospheric water vapour is enhanced.

In this proposed scenario, frost deposited at night rapidly sublimates in the morning and condenses in cold traps. Condensation in excess of a single molecular monolayer allows the H_2O molecules to behave as a liquid and mobilize ions in salts. Diurnal, or perhaps seasonal, cycling of these thin films of water over geologic timescales may be sufficient to concentrate Br to the observed levels.

The concentrations of Ni at Gusev are approximately 200 p.p.m. in the interiors of rocks, 550 p.p.m. in the dark soil, and 650 p.p.m. in the bright surface dust. Nickel is present in chondritic (CI) meteorites at an average level of 1.1% (ref. 45), and the difference between the Ni abundance in rocks and dust can be accounted for by adding 3.4% chondritic material to the rock composition. This approach, however, produces significant mismatches in other elements (Supplementary Fig. C). Thus, the surface dust at Gusev is not simply a product of meteoritic additions to a local rock composition. The 100 p.p.m. enhancement of Ni in bright surface dust relative to the dark soil does not necessarily result from, but is compatible with, a 1.2% addition of CI material (Supplementary Fig. C). This value is

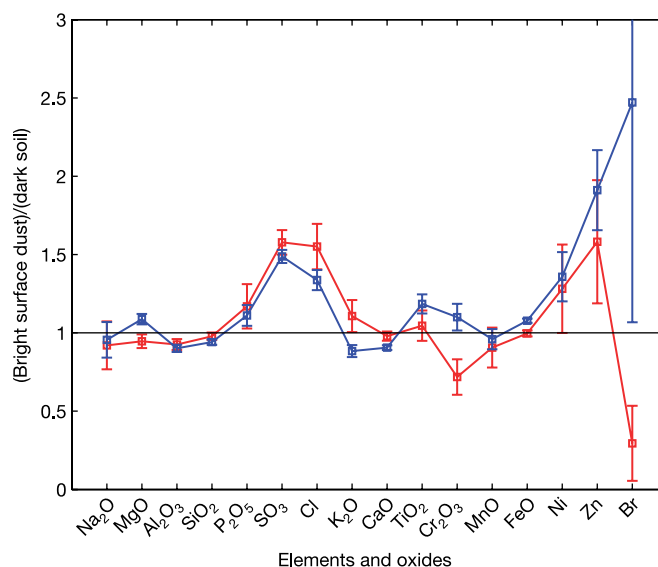


Figure 6 | Ratios of bright surface dust to dark soil exhibit similar trends for Gusev (red) and Meridiani (blue). Error bars represent 2-sigma statistical errors in the data.

consistent with predictions of a meteoritic component in martian soils^{46,47} and is comparable to the estimated admixture of 1.9% chondritic material in lunar soils⁴⁸.

Overview

The bright dust at the immediate surface of Mars is a globally distributed unit. The dark soils at Gusev and Meridiani are similar in composition and may also represent a distinct global unit, or given the apparent uniformity of basaltic terrains mapped from orbit, the connection between these dark soils may be a result of the general similarity in the rocks from which they originated. The fine-grained soil components at the MER sites are not derived from the local rocks and are products of wind redistribution. Oxidative weathering of the soil has not been extensive, suggesting rather limited interactions with liquid water. The action of thin films of water, possibly under current climatic conditions, is indicated by the distribution of bromine.

Received 20 November 2004; accepted 8 April 2005.

1. Binder, A. B. *et al.* The geology of the Viking Lander 1 Site. *J. Geophys. Res.* **82**, 4439–4451 (1977).
2. Mutch, T. A., Arvidson, R. E., Binder, A. B., Guinness, E. A. & Morris, E. C. The geology of the Viking 2 Site. *J. Geophys. Res.* **82**, 4452–4467 (1977).
3. Clark, B. C. *et al.* Chemical composition of the Martian Fines. *J. Geophys. Res.* **87**, 10059–10067 (1982).
4. Clark, B. C. *et al.* The Viking X ray fluorescence experiment: Analytical methods and early results. *J. Geophys. Res.* **82**, 4577–4594 (1977).
5. Golombek, M. P. *et al.* Overview of the Mars Pathfinder Mission and assessment of landing site predictions. *Science* **278**, 1743–1748 (1997).
6. Rieder, R., Wanke, H., Economou, T. & Turkevich, A. Determination of the chemical composition of Martian soil and rocks: The alpha proton X ray spectrometer. *J. Geophys. Res.* **102**, 4027–4044 (1997).
7. Brückner, J., Dreibus, G., Rieder, R. & Wanke, H. Refined data of Alpha Proton X-ray Spectrometer analyses of soils and rocks at the Mars Pathfinder site: Implications for surface chemistry. *J. Geophys. Res.* **108**, doi:10.1029/2003JE002060 (2003).
8. Foley, C. N., Economou, T. & Clayton, R. N. Final chemical results from the Mars Pathfinder alpha proton X-ray spectrometer. *J. Geophys. Res.* **108**, doi:10.1029/2003JE002019 (2003).
9. Bishop, J. L., Murchie, S. L., Pieters, C. M. & Zent, A. P. A model of dust, soil, and rock coatings on Mars: Physical and chemical processes on the Martian surface. *J. Geophys. Res.* **107**, doi:10.1029/2001JE001581 (2002).
10. Huck, F. O. *et al.* Spectrophotometric and color estimates of the Viking Lander sites. *J. Geophys. Res.* **82**, 4401–4411 (1977).

11. Smith, P. H. *et al.* Results from the Mars Pathfinder camera. *Science* **278**, 1758–1765 (1997).
12. Squyres, S. W. *et al.* The Spirit Rover's Athena Science Investigation at Gusev Crater, Mars. *Science* **305**, 794–799 (2004).
13. Squyres, S. W. *et al.* The Opportunity Rover's Athena Science Investigation at Meridiani Planum, Mars. *Science* **306**, 1698–1703 (2004).
14. Bell, J. F. III *et al.* Mars Exploration Rover Athena Panoramic Camera (Pancam) investigation. *J. Geophys. Res.* **108**, doi:10.1029/2003JE002070 (2003).
15. Christensen, P. R. *et al.* Miniature Thermal Emission Spectrometer for the Mars Exploration Rovers. *J. Geophys. Res.* **108**, doi:10.1029/2003JE002117 (2003).
16. Herkenhoff, K. E. *et al.* Athena Microscopic Imager investigation. *J. Geophys. Res.* **108**, doi:10.1029/2003JE002076 (2003).
17. Klingelhöfer, G. *et al.* Athena MIMOS II Mössbauer spectrometer investigation. *J. Geophys. Res.* **108**, doi:10.1029/2003JE002138 (2003).
18. Rieder, R. *et al.* The new Athena alpha particle X-ray spectrometer for the Mars Exploration Rovers. *J. Geophys. Res.* **108**, doi:10.1029/2003JE002150 (2003).
19. Gorevan, S. P. *et al.* Rock abrasion tool: Mars Exploration Rover mission. *J. Geophys. Res.* **108**, doi:10.1029/2003JE002061 (2003).
20. Madsen, M. B. *et al.* Magnetic properties experiments on the Mars Exploration Rover mission. *J. Geophys. Res.* **108**, doi:10.1029/2002JE002029 (2003).
21. Arvidson, R. E. *et al.* Localization and physical properties experiments conducted by Spirit at Gusev Crater. *Science* **305**, 821–824 (2004).
22. Klingelhöfer, G. *et al.* Jarosite and hematite at Meridiani Planum from Opportunity's Mössbauer spectrometer. *Science* **306**, 1740–1745 (2004).
23. Christensen, P. R. *et al.* Detection of crystalline hematite mineralization on Mars by the Thermal Emission Spectrometer: Evidence for near-surface water. *J. Geophys. Res.* **105**, 9623–9642 (2000).
24. Soderblom, L. A. *et al.* The soils of Eagle Crater and Meridiani Planum. *Science* **306**, 1723–1726 (2004).
25. Gellert, R. *et al.* Chemistry of rocks and soils in Gusev Crater from the Alpha Particle X-ray spectrometer. *Science* **305**, 829–832 (2004).
26. Morris, R. V. *et al.* Mineralogy at Gusev crater from the Mössbauer spectrometer on the Spirit rover. *Science* **305**, 833–836 (2004).
27. Lane, M. D., Dyar, M. D. & Bishop, J. L. Spectroscopic evidence for hydrous iron sulfate in the Martian soil. *Geophys. Res. Lett.* **31**, L19702, doi:10.1029/2004GL021231 (2004).
28. Bell, J. F. III *et al.* Mineralogic and compositional properties of martian soil and dust: results from Mars Pathfinder. *J. Geophys. Res.* **105**, 1721–1755 (2000).
29. Adams, J. B. & McCord, T. B. Mars: interpretation of spectral reflectivity of light and dark regions. *J. Geophys. Res.* **74**, 4851–4856 (1969).
30. Morris, R. V. *et al.* Mineralogy, composition, and alteration of Mars Pathfinder rocks and soils: Evidence from multispectral, elemental, and magnetic data on terrestrial analogue, SNC meteorite, and Pathfinder samples. *J. Geophys. Res.* **105**, 1757–1817 (2000).
31. Bandfield, J. L. & Smith, M. D. Multiple emission angle surface-atmosphere separations of Thermal Emission Spectrometer data. *Icarus* **161**, 47–65 (2003).
32. Christensen, P. R. *et al.* Initial results from the Mini-TES experiment in Gusev crater from the Spirit rover. *Science* **305**, 837–842 (2004).
33. Christensen, P. R. *et al.* Mineralogy at Meridiani Planum from the Mini-TES experiment on the Opportunity rover. *Science* **306**, 1733–1739 (2004).
34. Goetz, W. *et al.* Chemistry and mineralogy of magnetic dust at Gusev crater. *Nature* doi: 10.1038/nature03xxx (this issue).
35. Bertelsen, P. *et al.* Magnetic properties experiments on the Mars Exploration Rover Spirit at Gusev crater. *Science* **305**, 827–829 (2004).
36. Rieder, R. *et al.* Chemical composition of martian rocks and soils at Meridiani Planum from the Alpha Particle X-ray spectrometer. *Science* **306**, 1746–1749 (2004).
37. Bell, J. F. III *et al.* Pancam multispectral imaging results from the Opportunity rover at Meridiani Planum. *Science* **306**, 1703–1709 (2004).
38. Bell, J. F. III *et al.* Pancam multispectral imaging results from the Spirit rover at Gusev crater. *Science* **305**, 800–806 (2004).
39. Bandfield, J. L., Hamilton, V. E. & Christensen, P. R. A global view of Martian volcanic compositions. *Science* **287**, 1626–1630 (2000).
40. Christensen, P. R., Bandfield, J. L., Smith, M. D., Hamilton, V. E. & Clark, R. N. Identification of a basaltic component on the Martian surface from Thermal Emission Spectrometer data. *J. Geophys. Res.* **105**, 9609–9622 (2000).
41. Hamilton, V. E., Wyatt, M. B., McSween, H. Y. & Christensen, P. R. Analysis of terrestrial and martian volcanic compositions using thermal emission spectroscopy: II. Application to martian surface spectra from MGS TES. *J. Geophys. Res.* **106**, 14733–14747 (2001).
42. McSween, H. Y. *et al.* Basaltic rocks analyzed by the Spirit rover in Gusev crater. *Science* **305**, 842–845 (2004).
43. McLennan, S. M. Chemical composition of Martian soil and rocks: complex mixing and sedimentary transport. *Geophys. Res. Lett.* **27**, 1335–1338 (2000).
44. Clark, B. C. & Baird, A. K. Is the Martian lithosphere sulfur rich? *J. Geophys. Res.* **84**, 8395–8403 (1979).
45. Anders, E. & Grevesse, N. Abundances of the elements: meteoritic and solar. *Geochim. Cosmochim. Acta* **53**, 197–214 (1989).
46. Flynn, G. J. & McKay, D. S. An assessment of the meteoritic contribution to the martian soil. *J. Geophys. Res.* **95**, 14497–14509 (1990).
47. Yen, A. S. Composition and color of martian soil from oxidation of meteoritic material. *Lunar Planet. Sci. Conf. XXXII*, 1766 (2001).
48. Ganapathy, R., Keays, R. R., Laul, J. C. & Anders, E. *Proc. Apollo 11 Lunar Sci. Conf. Vol. 2 Chemical and Isotope Analyses* 1117–1142 (Pergamon, New York, 1970).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the members of the MER project who enable daily science observations at the Spirit and Opportunity landing sites. We thank J. Bishop and H. Newsom for providing reviews. The work described in this paper was conducted at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.Y. (Albert.Yen@jpl.nasa.gov).

Solar eclipses of Phobos and Deimos observed from the surface of Mars

J. F. Bell III¹, M. T. Lemmon², T. C. Duxbury³, M. Y. H. Hubbard¹, M. J. Wolff⁴, S. W. Squyres¹, L. Craig³ & J. M. Ludwinski³

The small martian satellites Phobos and Deimos orbit in synchronous rotation with inclinations of only 0.01° and 0.92° , respectively, relative to the planet's equatorial plane. Thus, an observer at near-equatorial latitudes on Mars could occasionally observe solar eclipses by these satellites (see ref. 1, for example). Because the apparent angular diameter of the satellites is much smaller than that of the Sun, however, such events are more appropriately referred to as transits. Transit data can be used for correcting and refining the orbital ephemerides of the moons. For example, Phobos is known to exhibit a secular acceleration that is caused by tidal dissipation within Mars^{2–4}. Long-term, accurate measurements are needed to refine the magnitude and origin of this dissipation within the martian interior as well as to refine the predicted orbital evolution of both satellites^{5,6}. Here we present observations of six transits of Phobos and Deimos across the solar disk from cameras on Mars aboard the Mars Exploration Rovers Spirit and Opportunity^{7,8}. These are the first direct imaging observations of satellites transiting the Sun from the surface of another planet.

Transits of Phobos were observed indirectly from Viking Lander camera light-curve imaging of the martian sky^{9,10} and from orbital imaging of the shadow cast by Phobos onto the planet^{4,11–13}. The new transit observations reported here were made with the Pancam imaging system⁸ carried by each rover. Pancam is a 1-cm-aperture, focal ratio $f/20$, $1,024 \times 1,024$ -pixel charge-coupled device camera capable of acquiring images of the Sun at a spatial scale of 0.28 mrad per pixel using neutral-density narrowband blocking filters. All transit observations were made using an 880 ± 20 -nm solar filter. We acquired images of four transits of Phobos and two transits of Deimos during March and April 2004 (Table 1). Four 'full' transits, two for Phobos and two for Deimos, were observed such that the standard contact point times could be derived; the other two observed transits were 'grazing' events for Phobos.

The first four events were observed from the Mars Exploration Rover MER-B rover Opportunity, which landed at 354.47417° E, 1.9483° S, and a radius of 3394.1482 km (Mars coordinates fixed by the International Astronomical Union, IAU, in 2000) and the last two events were observed from the MER-A rover Spirit, which landed at Mars coordinates of 175.4729° E, 14.5692° S, and a radius of 3392.2997 km (refs 14, 15). The positions of the rovers were modelled using the Jet Propulsion Laboratory (JPL) Navigation and Ancillary Information Facility (NAIF) spacecraft, planet, instruments, C-matrix, and events (SPICE) kernels mer1_ls_040128_iau2000_v1.bsp, mer1_surf_rover_0405181049.bsp, mer2_ls_040108_iau2000_v1.bsp, and mer2_surf_rover_0405172316.bsp (available at <ftp://naif.jpl.nasa.gov/pub/naif/MER/kernels/spk>). For each transit image acquired, rover flight software extracted a 63×63 -pixel image centred on the solar disk, which is ~ 22 pixels

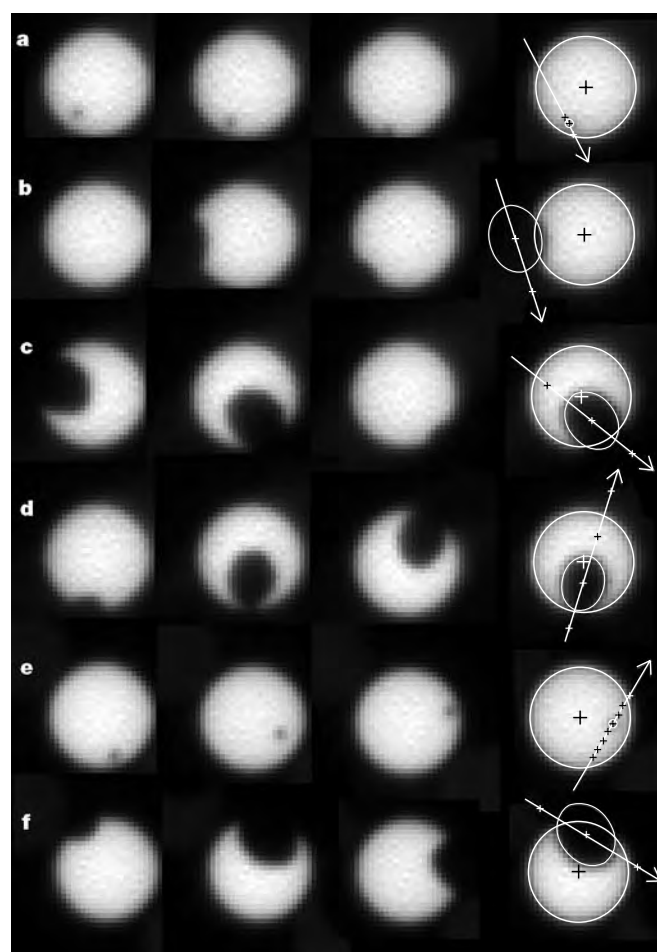


Figure 1 | Summary images of the transits observed by Spirit and Opportunity. Early contact or pre-contact images are shown in the first column, mid-transit images are in the second column, and late contact or post-contact images are shown in the third column. The fourth column shows our model of the path of the satellite during each event. **a**, MER-B Sol 39 Deimos transit of 2004-03-04. Images shown at 10 s spacing. The observations began near the middle of the event. **b**, MER-B Sol 42 Phobos grazing transit of 2004-03-07. Images shown at 10 s spacing. **c**, MER-B Sol 45 Phobos transit of 2004-03-10. Images shown at 10 s spacing. **d**, MER-B Sol 47 Phobos transit of 2004-03-12. Images shown at 10 s spacing. **e**, MER-A Sol 68 Deimos transit of 2004-03-13. Images shown at 30 s spacing. **f**, MER-A Sol 104 Phobos grazing transit of 2004-04-18. Images shown at 10 s spacing.

¹Cornell University, Ithaca, New York 14853-6801, USA. ²Texas A&M University, College Station, Texas 77843, USA. ³Jet Propulsion Laboratory/Caltech, Pasadena, California 91109, USA. ⁴Space Science Institute, Boulder, Colorado 80301, USA.

Table 1 | Summary of observed transit events

Event	Satellite	MER	Sol	Date (UTC)	d_{Sun} (arcsec)	α_{obs} ($^{\circ}$)	α_{pred} ($^{\circ}$)	Contact point times (UTC hh:mm:ss)				D_{obs} (s)	D_{pred} (s)	ΔT (s)
								I	II	III	IV			
A	Deimos	B	39	2004-03-04	1,241	0.11	0.16	03:03:02	03:03:10	03:04:25	03:04:33	91	66	-3
B	Phobos	B	42	2004-03-07	1,238	0.23	0.20	02:46:31	n/a	n/a	02:46:45	14	20	-3
C	Phobos	B	45	2004-03-10	1,235	0.04	0.06	07:36:32	07:36:43	07:36:49	07:37:01	29	31	-3
D	Phobos	B	47	2004-03-12	1,233	0.02	0.04	13:40:58	13:41:10	13:41:19	13:41:31	32	34	-3
E	Deimos	A	68	2004-03-13	1,232	0.11	0.15	00:04:24	00:04:31	00:05:39	00:05:47	83	68	-2
F	Phobos	A	104	2004-04-18	1,202	0.12	0.11	21:06:05	n/a	n/a	21:06:31	26	29	<9

'Sol' is the number of martian solar days (24 h 39 min 35.244 s) after the landing of each rover; d_{Sun} is the apparent angular diameter of the Sun in arcsec; α_{obs} is the observed phase angle of the transit, measured as the angular distance between the centre of the Sun and the mid-transit position of the satellite; α_{pred} is the predicted phase angle from currently-available ephemeris information; Contact point times are listed as hour:minute:second in Universal Time Coordinated (UTC); D_{obs} is the observed (modelled) duration of the transit in seconds; D_{pred} is the predicted duration from currently-available ephemeris information; ΔT is the estimated difference (due to clock drift) of the actual UTC time minus the UTC time as recorded on each rover, in sec. The uncertainty on ΔT is estimated to be between 2 to 5 s, but this is a systematic error that does not appear to introduce meaningful errors into our orbit-fitting procedure (see text).

across. Images were acquired every 10 s during each of the transit events, which is the fastest full-resolution frame rate possible for the rover imaging system. Because the satellites moved rapidly relative to the solar disk, the duration of the full transits were only 29 to 91 s and the partial transits lasted only 14 and 26 s, and so each event was captured in only two to eight images. Summary images of each of the events are presented in Fig. 1.

We used an image centroiding algorithm to determine the relative positions of the centre of the satellite and the centre of the solar disk for each image and used that information to generate a simple linear model of satellite position versus time. The model fit was then used to estimate the times of the four canonical transit contact points listed in Table 1. The Sun from Mars has an apparent angular diameter ranging from 1,160 to 1,400 arcsec because of the large eccentricity of the planet's orbit. At the time of the observed transit events, the Sun had an apparent angular diameter of $\sim 1,200$ arcsec (Table 1). Phobos, with a size of $27 \times 22 \times 19$ km and a distance from the martian surface of $\sim 5,984$ km, has an apparent angular diameter at zenith of about 700 arcsec, but this decreases by $\sim 32\%$ if the satellite is near the horizon. Phobos was observed to transit at angles ranging from 24° to 56° from the zenith, and was approximated in our model as an ellipse with a semimajor axis of 5.7 to 7.0 pixels and a semiminor axis of 4.4 to 5.4 pixels, depending on zenith angle. Deimos, with a size of $16 \times 12 \times 10$ km and a distance from the martian surface of $\sim 20,060$ km, has an apparent angular diameter at zenith of about 135 arcsec, decreasing by only $\sim 14\%$ with the satellite near the horizon. Because the two Deimos transits occurred at zenith angles of 22° and 33° , for this analysis the satellite was adequately

approximated as a circle with a constant diameter of 1.6 pixels. The solar disk was modelled as a circle with a diameter of 20 to 22 pixels. Contact points were defined as occurring when any part of the modelled Deimos circle or Phobos ellipse intersected the modelled solar circle. The phase angle was defined as the angular distance between the centre of the modelled solar disk and the closest point on the best-fit linear model of satellite position versus time. Because none of the transits were observed with the satellites close to the horizon, the effects of differential atmospheric refraction on the relative positions of the moons and sun were negligible at the 1-km-level accuracy of the results presented here.

The observed times of contact, corrected for systematic rover clock drift uncertainties, as well as the full images showing the moon positions during transit, were compared to the predicted times and images using the JPL Development Ephemeris mar033-7.bsp, which has not been updated since 1989 (refs 13, 16). Both local in-plane and out-of-plane orbit corrections were determined by matching simulated images to actual images (Fig. 1 shows the actual images and Fig. 2 shows an example of a simulated image).

Derived offsets between the predicted and observed positions of Phobos indicated that it was ahead of its predicted position by about 11 km ($+0.071^{\circ}$ in orbital longitude) and below by about 0.5 km (-0.003° out-of-plane). All 16 images analysed in the four Phobos events gave this same correction to the 1-km-level accuracy of the observations, providing confidence in the analysis methods and confirming that systematic effects like rover clock drift are being correctly incorporated in our model. The predicted Phobos position accuracy of the ephemeris is 10 km (1σ) at the times of the observations, in good agreement with the observed error. A similar comparison of the two Deimos events indicated that Deimos was about 38 km ($+0.125^{\circ}$ in orbital longitude) ahead of its predicted position and below by 25 km (-0.061° out-of-plane), also to the 1-km measurement accuracy level. The predicted Deimos position accuracy of the ephemeris was 20 km (1σ), also in reasonable agreement with the observed position correction.

The observed Phobos change is in the direction of increasing secular acceleration. If the observed longitude correction were attributed totally to orbital secular acceleration, then the current value of approximately $0.0013^{\circ}\text{yr}^{-2}$ would have to be increased by $0.0002^{\circ}\text{yr}^{-2}$ to match our observations. However, this small possible acceleration drift is comparable to the current uncertainty level of the orbital parameters. Indeed, within the uncertainties, the observed Phobos change could also be due to an error in the epoch orbital parameter values, or their secular rates (for example, mean anomaly or pericentre rate). Therefore, the variations observed in these new transit observations are not statistically significant enough to draw unique conclusions about potential variations in the secular acceleration value.

The limited number of observations, the small ranges in orbital longitude (less than 40° for the four Phobos transits and less than 20° for the two Deimos transits), and the fact that the observed satellite offsets barely exceed the 1σ errors from the propagated ephemerides,

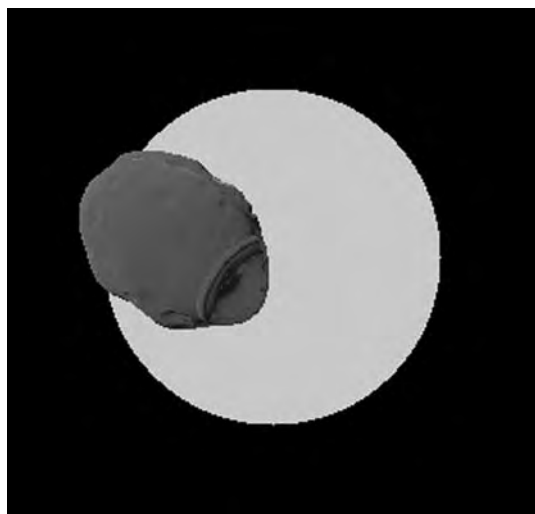


Figure 2 | Simulation of the position and orientation of Phobos relative to the Sun during the MER-B Sol 45 transit event of 2003-03-10 allowing both in-plane and out-of-plane orbit correction assessments.

prevent us from more uniquely constraining the current orbit parameters, including secular rates. However, if additional observations can be obtained, either from more orbital shadow imaging^{4,13} or possibly from the surface if the next transit seasons can be observed by the MER rovers in 2005 and/or 2006, then we will potentially be able to 'sense' the local epoch pericentre by having sufficient orbital longitude coverage, and then map the orbital longitude correction into mean anomaly and mean motion. The accuracy of the secular acceleration value would then be approximately 1 km (positional accuracy derived from transits) divided by ~ 17 yr squared since the last ephemeris update, or approximately $0.00002^\circ \text{yr}^{-2}$. Thus, under the best-case scenario of additional surface observations spanning a wider range of orbital longitude, it should be possible to expand these results to include significantly improved constraints on satellite secular accelerations.

Received 4 June 2004; accepted 31 January 2005.

- Ledger, E. The eclipses of the satellites of Mars. *The Observatory* **3**, 191–193 (1879).
- Smith, J. C. & Born, G. H. Secular acceleration of Phobos and Q of Mars. *Icarus* **27**, 51–53 (1976).
- Born, G. H. & Duxbury, T. C. The motions of Phobos and Deimos from Mariner 9 TV data. *Celest. Mech.* **12**, 77–88 (1975).
- Bills, B. G. *et al.* Improved estimate of tidal dissipation within Mars from MOLA observations of the shadow of Phobos. *J. Geophys. Res.* (in the press).
- Sinclair, A. T. The orbits of the satellites of Mars determined from Earth-based and spacecraft observations. *Astron. Astrophys.* **220**, 321–328 (1989).
- Mignard, F. Evolution of the Martian satellites. *Mon. Not. R. Astron. Soc.* **194**, 365–379 (1981).
- Squyres, S. W. *et al.* The Athena Mars Rover Science Investigation. *J. Geophys. Res.* **108**, 8062, doi:10.1029/2003JE002121 (2003).
- Bell, J. F. III *et al.* The Mars Exploration Rover Athena Panoramic Camera (Pancam) Investigation. *J. Geophys. Res.* **108**, 8063, doi:10.1029/2003JE002070 (2003).
- Viking Lander Imaging Team. *The Martian Landscape* NASA SP-425 (US Government Printing Office, Washington, DC, 1978).
- Bills, B. G. & Comstock, R. Solar eclipses on Mars: Spatial and temporal patterns in the motion of the shadow of Phobos. *Eos Trans. AGU* **84** (Fall Meeting Suppl.), abstr. G41B–30 (2003).
- Duxbury, T. C. Phobos transit of Mars as viewed by the Viking cameras. *Science* **199**, 1201–1202 (1978).
- Betts, B. H., Murray, B. C. & Svitek, T. Thermal inertias in the upper millimeters of the Martian surface derived using Phobos' shadow. *J. Geophys. Res.* **100**, 5285–5296 (1995).
- Neumann, G. A., Bills, B. G., Smith, D. E. & Zuber, M. T. Refinement of Phobos ephemeris using Mars Orbiter Laser Altimeter radiometry. *Lunar Planet. Sci. Conf. XXXV* abstr. 1820 (2004).
- Arvidson, R. E. *et al.* Localization and physical properties experiments conducted by Spirit at Gusev crater. *Science* **305**, 821–824 (2004).
- Arvidson, R. E. *et al.* Localization and physical properties experiments conducted by Opportunity at Meridiani Planum. *Science* **306**, 1730–1733 (2004).
- Jacobson, R. A., Synnott, S. P. & Campbell, J. K. The orbits of the satellites of Mars from spacecraft and earthbased observations. *Astron. Astrophys.* **225**, 548–554 (1989).

Acknowledgements These observations were made possible by the hard work of the engineers and scientists working with the Mars Exploration Rovers Spirit and Opportunity. We thank D. Limonadi, S. Spohn and J. Swift for their assistance with MER clock drift corrections. We also thank B. Bills for critical comments that improved this paper. This research is supported by grants and contracts from the NASA Mars Exploration Rover project.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.F.B. (jfb8@cornell.edu).

LETTERS MARS

Aeolian processes at the Mars Exploration Rover Meridiani Planum landing site

R. Sullivan¹, D. Banfield¹, J. F. Bell III¹, W. Calvin², D. Fike³, M. Golombek⁴, R. Greeley⁵, J. Grotzinger³, K. Herkenhoff⁶, D. Jerolmack³, M. Malin⁷, D. Ming⁸, L. A. Soderblom⁶, S. W. Squyres¹, S. Thompson⁵, W. A. Watters³, C. M. Weitz⁹ & A. Yen⁴

The martian surface is a natural laboratory for testing our understanding of the physics of aeolian (wind-related) processes in an environment different from that of Earth. Martian surface markings and atmospheric opacity are time-variable, indicating that fine particles at the surface are mobilized regularly by wind^{1–3}. Regolith (unconsolidated surface material) at the Mars Exploration Rover Opportunity's landing site has been affected greatly by wind, which has created and reoriented bedforms, sorted grains, and eroded bedrock. Aeolian features here preserve a unique record of changing wind direction and wind strength. Here we present an *in situ* examination of a martian bright wind streak, which provides evidence consistent with a previously proposed formational model^{4,5} for such features. We also show that a widely used criterion for distinguishing between aeolian saltation- and suspension-dominated grain behaviour is different on Mars, and that estimated wind friction speeds between 2 and 3 m s^{−1}, most recently from the northwest, are associated with recent global dust storms, providing ground truth for climate model predictions.

Pre-landing orbiter observations of Opportunity's landing site showed bright and dark streaks tapering away from craters on the Meridiani plains. From observations of similar features distributed across many locations on Mars, streak orientations indicate

formative wind directions^{6–8}. High-resolution images within the 117 km × 18 km landing ellipse obtained over several years show bright streak directions that indicate winds from the northwest and southeast⁹. What has not been recognized previously, however, is that this apparent bimodality of wind direction has a significant time dependence. Images obtained before the major 2001 dust storm are more likely to show bright streaks oriented in the opposite direction from images obtained after the storm waned. Rare overlapping image pairs provide two examples of an individual streak changing orientation after the intervening 2001 dust storm (Fig. 1). On this basis, we conclude that bright wind streak materials encountered by Opportunity are transient deposits mobilized by strong winds associated with major dust storms. Opportunity performed the first *in situ* investigation of a martian wind streak, evaluating the origin of Eagle crater's bright wind streak against predictive models^{4,5,10}. Opportunity focused on a bright patch of material just outside the southeast rim of Eagle crater. Pancam visible/near-infrared spectra of this material are similar to global air fall dust measured telescopically¹¹. Miniature Thermal Emission Spectrometer data of the same material can be modelled with ~70% air fall dust. Alpha Particle X-ray Spectrometer (APXS) and Mössbauer data also permit an air fall dust interpretation^{12–14}. The APXS results

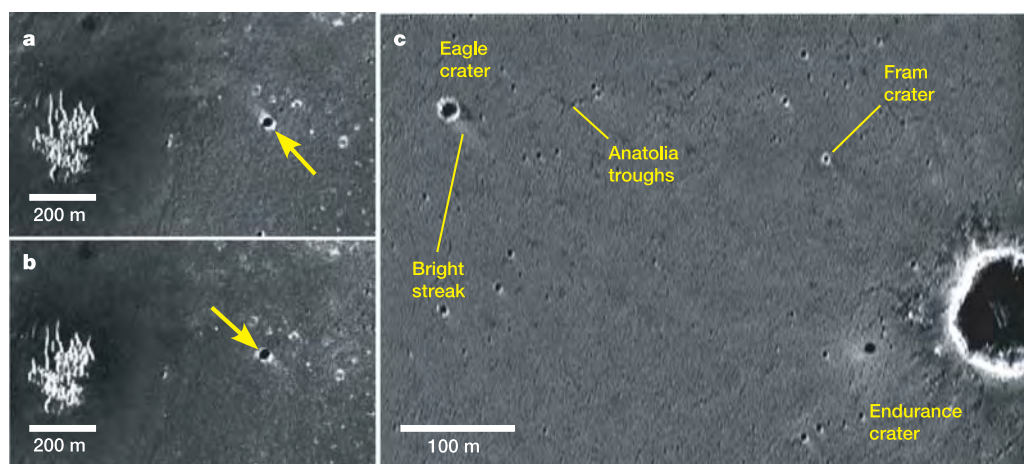


Figure 1 | Bright wind streaks on Meridiani Planum. **a, b**, Mars Global Surveyor Mars Orbiter Camera (MOC) views of a reversing bright wind streak at 5.3°W, 2.0°S. A global dust storm affected the planet during the

16-month interval between MOC images E0200373 (top) and E1800855 (bottom). **c**, Mars Exploration Rover descent camera image of the landing site, showing landmarks referred to in the text.

¹Department of Astronomy/CRSR, Cornell University, Ithaca, New York 14853, USA. ²Geological Sciences, University of Nevada, Reno, Nevada 89557, USA. ³Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ⁴Jet Propulsion Laboratory, Pasadena, California 90119, USA. ⁵Department of Geological Sciences, Arizona State University, Tempe, Arizona 85287-1404, USA. ⁶US Geological Survey, Flagstaff, Arizona 86001, USA. ⁷Malin Space Science Systems, San Diego, California 92191, USA. ⁸NASA Johnson Space Center, Houston, Texas 77058, USA. ⁹Planetary Science Institute, Tucson, Arizona 85719, USA.

show the bright material is not derived from the light-toned, sulphate-enriched outcrop within the crater. The streak itself derives its brightness from small bright deposits in areas associated with local surface roughness. We conclude that air fall dust, deposited in the partial wind shadow of Eagle crater, is responsible for the bright streak seen from orbit. These findings are consistent with a model proposed using relatively low-resolution Viking Orbiter data that predicted that patchy, discontinuous deposits of air fall dust would be distributed behind obstacles during periods of atmospheric thermal stability, for example, during major dust storms^{4,5}.

The wind friction speed u_* is $(\tau/\rho)^{1/2}$, where τ is shear stress applied to the surface and ρ is atmospheric density. u_{*t} is the threshold value of u_* required to initiate movement of a bed of particles, and is a function of gravity, and atmospheric and particle characteristics. The gradual transition from aeolian saltation to suspension is commonly defined where wind friction speed, u_* , for a particle begins to exceed its terminal fall velocity, u_F (refs 15–18). For terrestrial conditions this particle size is around 50 μm , but is expected to be around 200 μm for Mars owing to higher u_{*t} values for all particle sizes and lower gravity^{16,17}. On Earth, dune sand is several times the transitional particle size, so this is anticipated also to be the case on Mars¹⁹. However, ripples on the floor of Eagle crater are

composed of basaltic grains 50–125 μm in size. The u_F/u_{*t} ratio for these grains is only ~ 0.5 . Therefore, the relationship between the u_F/u_{*t} ratio and particle behaviour must be different for Mars.

Under terrestrial conditions, ripples of silt grains can form with $u_F/u_{*t} < 1$, but these bedforms are in general confined to wind tunnels, under conditions of u_* only slightly exceeding u_{*t} , so that mobilized particles remain concentrated closely above the bed and particle trajectories are still relatively short²⁰. It seems unlikely that these specific conditions, scaled appropriately for Mars, are indicated by the Eagle floor ripples composed of 50–125- μm grains and other deposits in temporary aeolian traps seen along Opportunity's traverse. Another, less-restrictive explanation might involve the potential inadequacy of u_F/u_{*t} to characterize particle responsiveness to turbulence. A particle small enough to be lofted into suspension must be more responsive to the accelerations of turbulent eddies than to gravity. Therefore particle response time to achieve u_F might be significant. The lower martian atmospheric density (by a factor of 80) and $\sim 25\%$ lower atmospheric dynamic viscosity should lengthen the particle response time to accelerations of turbulent eddies, requiring increased turbulent energy from higher wind speeds to achieve the same probability of suspension as on Earth. In this scenario, the transition from saltation to suspension on Mars

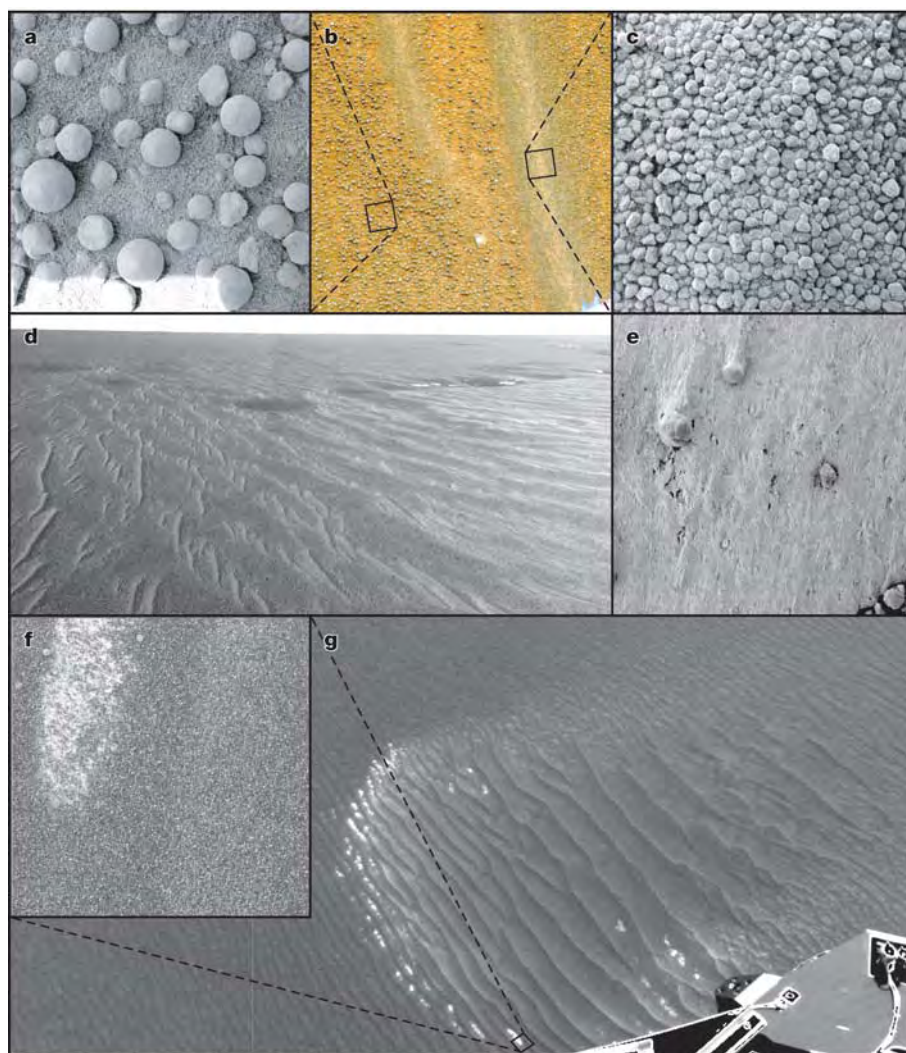


Figure 2 | Bedforms and ventifacts at Meridiani Planum. Microscopic Imager views are ~ 31 mm across. **a–c**, False-colour Pancam (753 nm, 535 nm, 440 nm) view showing locations of Microscopic Imager close-ups of a ripple crest and an area between ripples. **d**, Navcam mosaic spanning 80° showing plains ripples organized into primary and secondary orientations.

e, Ventifacts on sulphate-enriched outcrop at Fram crater. **f**, Microscopic Imager view of basaltic sand ripples on Eagle crater floor. **g**, Navcam view of the same basaltic sand ripples. The longest wavelengths are ~ 11 cm. Brighter material in some ripple troughs has been distributed to the right, downwind to the southeast.

is delayed to greater turbulent energies and smaller particle sizes than the u_F/u_{*t} ratio would suggest. Some support for this concept is provided by numerical simulations of martian saltation particle trajectories that predict that sand grains on Mars should achieve smaller fractions of entraining wind speed along their saltation trajectories than on Earth²¹.

Bedforms across the site can be organized into two groups based on correlated differences in constituents, morphology, and local setting: (1) plains ripples armoured with coarse sand; and (2) patches of ripples composed of 50–125- μm basaltic sand that occur in local depressions. The wind direction indicated for basaltic sand ripples is aligned with the current bright streak direction where these features occur together, but is not aligned with winds indicated for the more extensive plains ripples. On the basis of these differences and the different wind speeds required to mobilize each bedform type, we develop a brief, recent aeolian history of the site that constrains estimates of local, near-surface wind speed and direction during recent major dust storms.

Ripples composed almost entirely of basaltic 50–125- μm sand are found within craters, smaller pits, and structural troughs that apparently serve as particle traps (Fig. 2). Ripple profiles have smooth, continuously concave troughs connected by narrow crests (different from plains ripples). These sand ripples were examined on the floor of Eagle crater and on drifts among the Eagle crater outcrops, and were observed more distantly within the Anatolia trough system, the 9-m-diameter crater Fram, and within Endurance crater. The Eagle floor ripples occur in several distinct patches with different characteristic wavelengths. Orientations are about N38°E, and bright material deposited within some ripple troughs appears to have been partially redistributed approximately southeast along the crater floor, consistent with the most recent winds forming the much-larger exterior bright wind streak that extends S45°E from the crater. The Eagle crater floor ripples were trenched more easily by the rover wheel than any other material²², and open pore space between sand grains in these ripples is apparent in Microscopic Imager views. These bedforms are probably currently mobilized by the same recent winds that formed the bright wind streaks associated with Eagle crater and other craters in the area.

Plains ripples (Fig. 2) are covered with a monolayer of rounded, 1–2-mm fragments of haematite-enriched concretions, and have poorly sorted interiors dominated by a matrix of fine to very fine sand. Feature heights and widths are typically ~ 1 cm and 10 cm, respectively, and crest-line lengths range up to 2 m. Larger examples occur on rims adjacent to shallow depressions. Areas between these ripples are dominated by 50–125- μm basaltic grains that partly bury additional concretion fragments. Intact, spherical concretions several millimetres across are found scattered only between ripples²³. Mössbauer contact plate impressions into areas between plains ripples reveal weak cohesion between basaltic grains. Unlike the loose basaltic sand grains observed within Eagle and Endurance craters, much of the 50–125- μm sand out on the plains has evidently remained inactive long enough for cohesion to form between grains by some unknown process. This implies similar inactivity for smaller and larger particles there (which are moved by wind less easily) and the plains ripples they compose. We conclude that plains ripples have not been active as recently as other bedforms described above. Individual plains ripples are oriented about $\sim \text{N}26^\circ\text{E}$ but are commonly grouped in echelon in alignments defining an older orientation of about $\text{N}4^\circ\text{E}$. A clockwise change in wind direction is indicated from the older to the younger orientation of plains ripples, and then to the more recently active basalt sand ripples within Eagle (N38°E) and the Eagle crater bright wind streak (Fig. 3).

Ventifacts (erosional rock features formed by abrasion from wind-blown particles) provide a cumulative record of aeolian activity consistent with the changing wind patterns indicated by bedforms and albedo features. Small, elongate protrusions of light-toned rock occur adjacent to a subset of partially embedded, haematite-enriched

concretions. These features ('rock tails') are interpreted as erosional remnants in wind shadows behind the more resistant concretions (Fig. 2). Because these features occur in rock, they probably record an integrated, long-timescale history of the strongest, rarest wind events, and possibly the directions to major sediment sources. Concretion rock tail morphologies include raised ridges that taper in height and width away from concretions, as well as pedestals supporting concretions at their tips. Rock tails do not occur in every rock unit exposed at Eagle and Endurance craters. Orientations of rock tails show two lobes of formative winds: from $284 \pm 28^\circ$ ($n = 107$), and from $135 \pm 25^\circ$ ($n = 25$). The major lobe is aligned with winds from the west driving plains ripples (the $\text{N}26^\circ\text{E}$ and older $\text{N}4^\circ\text{E}$ orientations), and also includes some data consistent with the most recent strong winds from the northwest associated with the Eagle floor basaltic sand ripples and the bright wind streak. The second, minor lobe is consistent with winds from the southwest, which is the other recent wind direction associated with bright streaks in the landing ellipse.

A fundamental assumption is that the highest-energy wind events are the rarest, but can effect the most change. The preserved record is likely to reflect a biased series of progressively stronger and older but more obscure events, with even older or interleaved weaker events being unrepresented even if they were more typical. The most recent events at the landing site involve cycles of bright wind streak erasure and formation, caused by winds from approximately 315° or 135° exceeding $u_{*t} \approx 2.0 \text{ ms}^{-1}$ (corresponding to a wind speed u of $\sim 45 \text{ m s}^{-1}$ at 1 m above the bed) associated with large dust storms (Fig. 3). These wind events activate basaltic sand ripples in temporary particle traps like the floor of Eagle crater and, when strong enough, sweep accumulations of trapped material out on to the plains, supplying a sparse population of mobile sand grains that migrates from one trap to another. The armoured $\text{N}26^\circ\text{E}$ plains ripples have

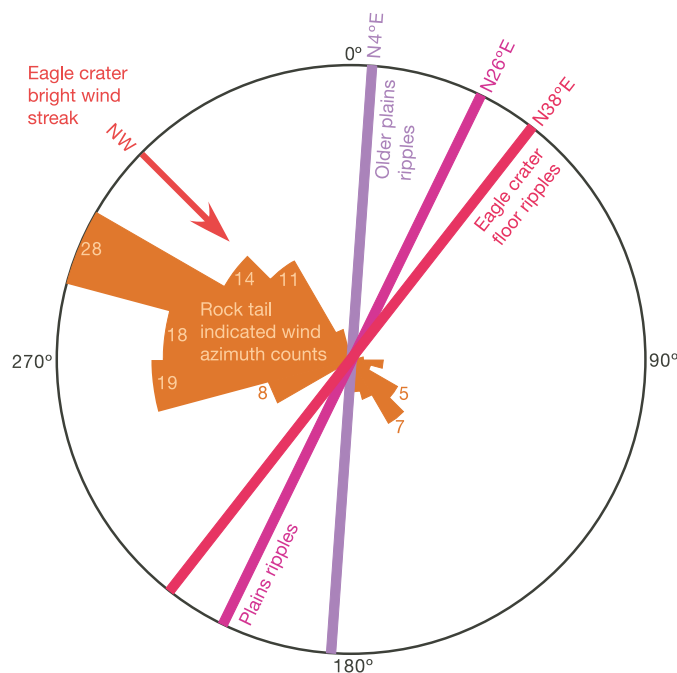


Figure 3 | Orientations of rock tails, plains ripples, basaltic sand ripples on the floor of Eagle crater, and the Eagle crater bright wind streak. Rock tail orientations (brown, with larger counts noted) span an azimuth range consistent with winds driving plains ripples (current and former orientations) in current $\text{N}26^\circ\text{E}$ and former $\text{N}4^\circ\text{E}$ orientations. Rock tails also seem to record erosion from winds consistent with the current wind streak direction and its opposite. A 40° clockwise rotation in major wind axes is indicated.

not been active as recently. These ripples formed when one or more strong wind events from around 296° partially reoriented a pre-existing set of plains bedforms. Winds during this process probably moved the 1–2-mm concretion fragments in creep only, driven by basaltic sand. Stronger winds capable of saltating the concretion fragments directly—exceeding $u_* \approx 3.0 \text{ m s}^{-1}$ ($u \approx 70 \text{ m s}^{-1}$ at 1 m)—probably would not have preserved any sign of the pre-existing N4°E bedform orientation.

However, excellent sorting of the 1–2-mm concretion fragment population currently armouring plains ripples indicates that very strong winds in the more distant past were capable of saltating these particles. These winds, with $u_* > 3.0 \text{ m s}^{-1}$ ($u > 70 \text{ m s}^{-1}$ at 1 m) may have been responsible for fragmenting a source supply of larger concretions through energetic collisions during transport, as well as sorting the fragments. It is unknown how the sorted population of concretion fragments armouring plains ripples became intermingled with the larger spherical concretions distributed across the flat areas between plains ripples. Winds capable of winnowing away finer grains but not coarser clasts can form deflational lags. As the surface concentration of larger, immobile clasts increases, the fraction of wind shear stress available to move loose, finer grains in between is reduced. Experiments with particle size frequencies broadly similar to inter-ripple areas at the landing site^{24,25} suggest u_{*t} to move the finer grains will increase to speeds similar to those required to saltate the concretion fragments covering plains ripples. Thus deflation seems to have modified the plains to a uniform threshold shear velocity condition, with further deflation slowed by concretion fragment armour on the ripples, and larger, intact spherical concretions scattered across the flat areas between. The present-day inactivity of plains ripples indicates an upper bound for local u_* associated with current dust storms of $\sim 3 \text{ m s}^{-1}$, up to $\sim 4 \text{ m s}^{-1}$ if the concretion fragments are pure haematite.

METHODS

Calculations of u_{*t} use expressions derived from wind-tunnel experiments^{26,27} and we specify the following conditions for all calculations except where noted: 250°K, 7 mbar CO₂, martian gravity, and 3.0 g cm^{-3} particles. Wind speed, u , at height 1 m was estimated using an aerodynamic roughness of $1 \times 10^{-4} \text{ m}$ in a standard logarithmic wind profile. This value of z_0 is representative, used for comparative purposes only; for a given u_* or u_{*t} value, variations in z_0 from $5 \times 10^{-5} \text{ m}$ to $5 \times 10^{-4} \text{ m}$ will raise or lower, respectively, all u estimates at $z = 1 \text{ m}$ by 7–17%.

The lower bound of the regolith particle size range is uncertain because it is below the resolution (30 μm per pixel) of the Microscopic Imager²⁸. This is especially relevant for regolith at the depth revealed by wheel trenching, which is cloddy and contains substantial fractions of grains unresolved by the Microscopic Imager. However, the fraction of basaltic grains $< 50 \mu\text{m}$ at the surface is likely to be small, because most grains on bulk surface deposits are resolved (2–5 pixels), as are individual dark grains silhouetted on light-toned outcrop²³.

Received 23 November 2004; accepted 12 April 2005.

- Martin, L. J. & Zurek, R. W. An analysis of the history of dust activity on Mars. *J. Geophys. Res.* **98**, 3221–3246 (1993).
- Greeley, R., Lancaster, N., Lee, S. & Thomas, P. in *Mars* (eds Kieffer, H., Jakosky, B., Snyder, C. & Matthews, M. U.) 730–766 (Arizona Press, Tucson, 1992).
- Cantor, B. A., James, P. B., Caplinger, M. & Wolfe, M. J. Martian dust storms: 1999 Mars Orbiter Camera observations. *J. Geophys. Res.* **106**, 23653–23687 (2001).
- Veverka, J., Gierasch, P. & Thomas, P. Wind streaks on Mars: meteorological control of occurrence and mode of formation. *Icarus* **45**, 154–166 (1981).
- Thomas, P., Veverka, J., Gineris, D. & Wong, L. Dust streaks on Mars. *Icarus* **60**, 161–179 (1984).

- Sagan, C. *et al.* Variable features on Mars. 2. Mariner 9 global results. *J. Geophys. Res.* **78**, 4163–4196 (1973).
- Arvidson, R. E. Wind-blown streaks, splotches, and associated craters on Mars: Statistical analysis of Mariner 9 photographs. *Icarus* **21**, 12–27 (1974).
- Thomas, P. C., Veverka, J., Lee, S. & Bloom, A. Classification of wind streaks on Mars. *Icarus* **45**, 124–153 (1981).
- Greeley, R. & Thompson, S. D. Mars: Aeolian features and wind predictions at the Terra Meridiani and Isidis Planitia potential Mars Exploration Rover landing sites. *J. Geophys. Res.* **108**(E12), doi:10.1029/2003JE002110 (2003).
- Greeley, R. Wind tunnel simulations of light and dark streaks on Mars. *Science* **183**, 847–849 (1974).
- Bell, J. F. III, McCord, T. B. & Owensby, P. D. Observational evidence of crystalline iron oxides on Mars. *J. Geophys. Res.* **95**, 14447–14461 (1990).
- Klingelhöfer, G. *et al.* Jarosite and hematite at Meridiani Planum from Opportunity's Mössbauer spectrometer. *Science* **306**, 1740–1745 (2004).
- Rieder, R. *et al.* Chemistry of rocks and soils at Meridiani Planum from the Alpha Particle X-ray Spectrometer. *Science* **306**, 1746–1749 (2004).
- Yen, A. *et al.* An integrated view of the chemistry and mineralogy of martian soils. *Nature* doi:10.1038/nature03637 (this issue).
- Sagan, C. & Bagnold, R. A. Fluid transport on Earth and aeolian transport on Mars. *Icarus* **26**, 209–218 (1975).
- Iversen, J. D., Greeley, R. & Pollack, J. B. Windblown dust on Earth, Mars, and Venus. *J. Atmos. Sci.* **33**, 2425–2429 (1976).
- Greeley, R. & Iversen, J. D. *Wind as a Geological Process* 68–71 (Cambridge Univ. Press, Oxford, 1985).
- Tsoar, H. & Pye, K. Dust transport and the question of desert loess formation. *Sedimentology* **34**, 139–153 (1987).
- Edgett, K. S. & Christensen, P. R. The particle size of martian aeolian dunes. *J. Geophys. Res.* **96**, 22765–22776 (1991).
- Goossens, D. Aeolian dust ripples: Their occurrence, morphometrical characteristics, dynamics and origin. *Catena* **18**, 379–407 (1991).
- Greeley, R. *et al.* Rate of wind abrasion on Mars. *J. Geophys. Res.* **87**, 10009–10024 (1982).
- Arvidson, R. E. *et al.* Localization and physical properties experiments conducted by Opportunity at Meridiani Planum. *Science* **306**, 1730–1733 (2004).
- Soderblom, L. A. *et al.* Soils of Eagle Crater and Meridiani Planum at the Opportunity rover landing site. *Science* **306**, 1723–1726 (2004).
- Gillette, D. A. & Stockton, P. H. The effect of nonerodible particles on wind erosion of erodible surfaces. *J. Geophys. Res.* **94**, 12885–12893 (1989).
- Nickling, W. G. & McKenna Neuman, C. Development of deflation lag surfaces. *Sedimentology* **42**, 403–414 (1995).
- Iversen, J. D. & White, B. R. Saltation threshold on Earth, Mars and Venus. *Sedimentology* **29**, 111–119 (1982).
- White, B. R., Lacchia, B. M., Greeley, R. & Leach, R. N. Aeolian behaviour of dust in a simulated Martian environment. *J. Geophys. Res.* **102**, 25629–25640 (1997).
- Herkenhoff, K. *et al.* Evidence from Opportunity's Microscopic Imager for water on Meridiani Planum. *Science* **306**, 1727–1730 (2004).

Acknowledgements This work was performed for the Jet Propulsion Laboratory, California Institute of Technology, sponsored by the National Aeronautics and Space Administration. We are grateful for the efforts of the Mars Exploration Rover development and operations teams that made this work possible. We acknowledge the use of Mars Orbiter Camera images processed by Malin Space Science Systems that are available at http://www.msss.com/moc_gallery/.

Authors Contributions J.F.B. and W.C. provided Pancam and MiniTES analyses, respectively, of bright streak material. W.A.W. measured all rock tails. D.F. discovered the time dependence of wind streak orientations in MOC images. D.M. reviewed the APXS linear mixing work of R.S. R.S. also measured the ripple orientations, calculated u_{*t} and u values, worked out the aeolian history of the site, identified the discrepancy between particle size of basaltic ripples and u_F/u_{*t} ratio, and drafted the original and revised manuscripts. D.J. contributed key points relating to deflation at the site. R.S., D.J. and D.B. worked on potential explanations for the low u_F/u_{*t} ratio mobility of the basaltic sand. L.A.S. led the Science Operations Working Group during the bright streak rover operations. A.Y. advised on APXS calibration issues. All authors, particularly M.M., provided significant scientific guidance and/or editorial inputs.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.S. (rjs33@cornell.edu).

LETTERS MARS

Indication of drier periods on Mars from the chemistry and mineralogy of atmospheric dust

Walter Goetz¹, Preben Bertelsen², Charlotte S. Binau², Haraldur P. Gunnlaugsson³, Stubbe F. Hviid¹, Kjartan M. Kinch³, Daniel E. Madsen², Morten B. Madsen², Malte Olsen², Ralf Gellert^{4,5}, Göstar Klingelhöfer⁵, Douglas W. Ming⁶, Richard V. Morris⁶, Rudolf Rieder⁴, Daniel S. Rodionov^{5,7}, Paulo A. de Souza Jr^{5,8}, Christian Schröder⁵, Steve W. Squyres⁹, Tom Wdowiak¹⁰ & Albert Yen¹¹

The ubiquitous atmospheric dust on Mars is well mixed by periodic global dust storms, and such dust carries information about the environment in which it once formed and hence about the history of water on Mars¹. The Mars Exploration Rovers have permanent magnets to collect atmospheric dust for investigation by instruments on the rovers^{2,3}. Here we report results from Mössbauer spectroscopy and X-ray fluorescence of dust particles captured from the martian atmosphere by the magnets. The dust on the magnets contains magnetite and olivine; this indicates a basaltic origin of the dust and shows that magnetite, not maghemite, is the mineral mainly responsible for the magnetic properties of the dust. Furthermore, the dust on the magnets contains some ferric oxides, probably including nanocrystalline phases, so some alteration or oxidation of the basaltic dust seems to have occurred. The presence of olivine indicates that liquid water did not play a dominant role in the processes that formed the atmospheric dust.

The reddish dust particles suspended in the martian atmosphere have a mean diameter of less than 5 µm (refs 4–6). Airborne particles are collected on the magnets on board the Mars Exploration Rovers^{2,3,7,8}. Three types of magnets will be discussed here: the filter and capture magnets, and the sweep magnet. The sweep magnet is designed to allow only non-magnetic particles to enter into an area within the centre of a cylindrical magnet. The field from this magnet is so strong that it deflects the paths of all particles with any significant magnetic susceptibility³. The filter and capture magnets are designed to capture a layer of dust for investigation by the panoramic camera (Pancam) and the instruments on the robotic arm of the rovers. The capture magnet is very strong and will capture airborne particles with a wide range of magnetic properties, whereas the filter magnet will preferentially capture strongly magnetic particles³.

Throughout the mission, Pancam has been used to monitor the dust on each magnet. Occasionally the dust on the capture and filter magnets are investigated using the α -particle X-ray spectrometer (APXS), the Mössbauer spectrometer, and the Microscopic Imager. Investigation of particles collected by magnets on Mars during previous missions has shown that most of the airborne particles must contain a ferrimagnetic mineral⁹. Furthermore, on the Mars Exploration Rovers it has been possible to show unambiguously that the dust particles have a wide range of magnetic properties and therefore can be sorted accordingly: the filter magnet has assembled a more-magnetic and darker fraction of the dust particles than the

capture magnet¹⁰. A similar correlation between magnetic properties and colour was seen in the details of the dust pattern on the capture magnet¹¹. Imaging of the sweep magnet³ has shown that most—if not all—of the airborne dust particles are magnetic¹⁰. It is impossible to detect any particles in the centre of the sweep magnet in the Pancam images—presumably because the magnetic force has prevented particles from settling there.

During the mission, the APXS has been deployed on a series of different soils^{12,13}. The overall chemical composition of these soils is rather uniform and indicates a basaltic origin. Variability across the soil units at Gusev crater and Meridiani Planum is discussed elsewhere¹⁴. The thin layer of dust on the magnet aluminium surface³ complicates the quantitative analysis of the APXS spectra. The surface is only partially covered by dust, and the thickness of the layer is comparable to the penetration depth of characteristic X-rays from the major elements. A direct comparison of raw spectra would lead to a relative underestimation of heavier elements in the dust on the magnets compared to an infinitely thick layer. Nevertheless, firm general conclusions can be drawn from the APXS results. To a first approximation, the dust on the magnets has an elemental composition that is similar to average bright soil (Table 1). All major element peaks present in rocks and soils are also present in the dust on the magnets. By comparing line areas from the raw spectra for neighbouring elements, the complications associated with the different (energy-dependent) effective thicknesses for different elements

Table 1 | Compositions of martian soil and dust

Element	Soil	Dust A	Dust B
Na	0.035(2)	0.20(1)	0.116(5)
Mg	0.172(4)	0.53(2)	0.348(9)
Al*	0.266(5)	4.7(4)	13.1(9)
Si	1.00	1.00	1.00
P	0.027(2)	0.040(6)	0.037(3)
S	0.088(3)	0.109(7)	0.114(5)
Cl	0.021(1)	0.032(5)	0.037(3)
K	0.012(1)	0.013(4)	0.012(2)
Ca	0.094(3)	0.062(6)	0.071(4)
Ti	0.009(1)	0.009(4)	0.009(2)
Fe	0.355(5)	0.083(7)	0.100(5)

Areas of element peaks in APX spectra of soil (Spirit sol 14), and of dust on the capture magnet on the two rovers: A (Spirit sol 150) and B (Opportunity sol 168). The error on the last digit is given in parentheses. In each spectrum, the areas are normalized relative to Si. * The aluminium signal is mainly due to the high-purity aluminium plate on the magnet (below the dust).

¹Max-Planck-Institut für Sonnensystemforschung, Katlenburg-Lindau, D-37191, Germany. ²Niels Bohr Institute, University of Copenhagen, Copenhagen, DK-2100, Denmark.

³Institute of Physics and Astronomy, University of Aarhus, Aarhus, DK-8000, Denmark. ⁴Max-Planck-Institut für Chemie, Mainz, D-55128, Germany. ⁵Institut für Anorganische und Analytische Chemie, J. Gutenberg Universität, Mainz, D-55128, Germany. ⁶NASA Johnson Space Center, Houston, Texas 77058, USA. ⁷Space Research Institute IKI, Moscow, Russia. ⁸Companhia Vale do Rio Doce, 29090-900 Vitória, ES, Brazil. ⁹Cornell University, Ithaca, New York 14853, USA. ¹⁰University of Alabama, Birmingham, Alabama 35203, USA. ¹¹Jet Propulsion Laboratory—California Institute of Technology, Pasadena, California 91109, USA.

are partly avoided. As an example, we have selected the two pairs of elements Na and Mg, and P and Si, and plotted the relative peak areas of these pairs for rocks, soil and the dust present on the magnets (Fig. 1). Figure 1 shows that a classification of samples can be made on the basis of the ratios plotted. The figure is consistent with the hypothesis that the soil is composed of a rock component (of composition as Adirondack and Humphrey) and a dust component, which is enriched in Na and P (among other elements). This suggests that the bright soil is derived from a combination of basaltic sand (originating from the rocks) and a dust component similar to the dust on the magnets.

Figure 2 shows a Mössbauer spectrum of the dust attracted to the capture magnet on Opportunity. The spectrum was recorded from martian day (sol) 328 to sol 330, and over a total time interval of about 50 h. For comparison, the spectrum of a typical soil is also included in the figure.

The dust layer is of the order of 1 mg cm^{-2} thick, and the sampling depth of the Mössbauer spectrometer¹⁵ is of the order of 30 mg cm^{-2} . The intensity (and quality) of the spectrum is therefore low. From the striking similarity between the soil spectrum and the dust spectrum at velocities from -2 mm s^{-1} to $+4 \text{ mm s}^{-1}$, where paramagnetic compounds dominate, it is evident that these spectra contain similar paramagnetic components¹⁶. As for typical bright soils, these doublets are assigned to the phases olivine, pyroxene and paramagnetic and/or superparamagnetic Fe^{3+} -compounds, possibly including oxyhydroxides and haematite^{16–18}. It should be noted that the ferric doublet is more intense in the dust on the magnet than in typical soils.

One could argue that olivine and pyroxene are not unambiguous interpretations of the ferrous doublet components in this spectrum of dust on the magnets. However, the martian surface is known to act as a dynamic source and sink for airborne dust (for example, via dust tracks, dust devils, and local storms), which imposes some relationship between dust and soil mineralogy. Furthermore, given the similarity of soil and dust spectra (both Mössbauer and APXS) and the arguments that have led to the assignment of these minerals for the components of the soil spectra, it would seem somewhat far-fetched to assume that a different set of minerals with similar spectra (and in similar relative abundance) should be the cause of these components in the dust.

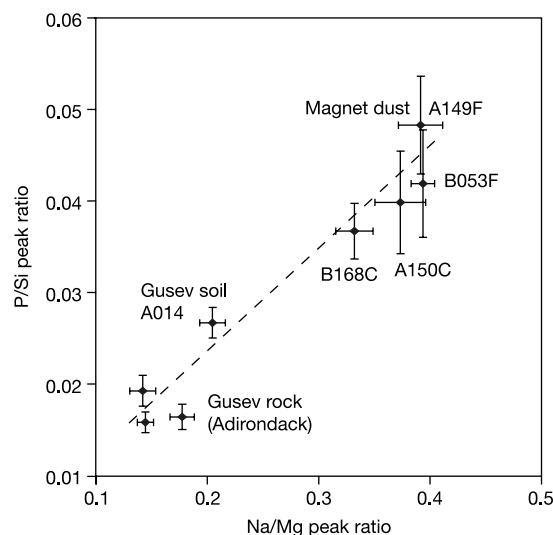


Figure 1 | Selected peak area ratios from APX spectra. Shown are data for rock and soil samples from Gusev crater, together with samples of dust from magnets labelled according to spacecraft (A, Spirit; B, Opportunity), measurement sol, and type of magnet (C, capture; F, filter). Error bars represent one standard deviation.

The magnetically ordered components in the spectrum are characterized by the following features: the scattered intensity in the velocity ranges around -8 mm s^{-1} and $+9 \text{ mm s}^{-1}$ show a broad scattering peak at -8 mm s^{-1} and a more narrow and intense peak at $+8.5 \text{ mm s}^{-1}$ —this pattern is indicative of the mineral magnetite (Fe_3O_4 ; see Methods). For comparison, we have included in Fig. 2 a Mössbauer spectrum of the compositional calibration target (CCT) of the rovers (this is a thin rock slab that is mounted under the rover solar panels for calibration of the instruments on the robotic arm). The CCT has a very high concentration of nearly stoichiometric magnetite. Comparing the dust spectrum with the spectrum of the CCT, it is clear that—to a first approximation—most components of the spectrum of the dust on the capture magnet can be accounted for by a sum of the spectra of the bright soil and of the magnetite of the CCT. This shows that the dust is magnetic because it contains the mineral magnetite (approximately 50% of the spectral area), and also that the dust contains major paramagnetic components. We note that in 1979, Pollack *et al.*⁶ suggested the presence of magnetite in martian dust, on the basis of the optical properties of the dust as observed by the Viking cameras.

Fitting of the spectrum shows that either the magnetite in the dust is not perfectly stoichiometric magnetite (the line at -7 mm s^{-1} is of significantly lower intensity than in the CCT), or that some ferric oxide (maghemite and/or haematite) is also present (see Methods for details).

On the basis of the Mössbauer spectrum of the dust on the Opportunity capture magnet (Fig. 2), it is clear that the most probable candidate for the ferrimagnetic mineral in the airborne dust is magnetite. It is possible that the ferrimagnetic mineral is a somewhat oxidized variant of magnetite—perhaps non-stoichiometric magnetite, or a solid solution of magnetite and maghemite. But neither maghemite alone (nor maghemite and haematite) can fit this spectrum adequately. The magnetite may also be isomorphically substituted with titanium. In the APX spectrum of the dust on the Opportunity capture magnet (not shown), the peak areas of all elements continuously increase during the first 150 sols. After a dust removal event (strong wind gust or dust devil around sol 200), we see

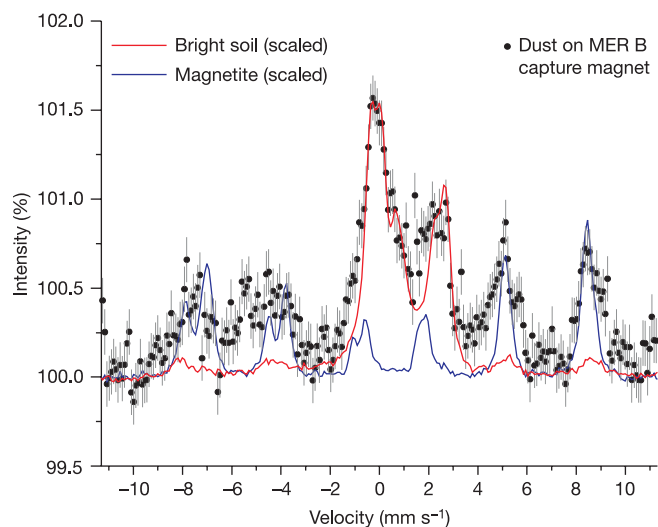


Figure 2 | Mössbauer spectrum of airborne dust. Mössbauer spectrum of the dust attracted to the capture magnet on Opportunity, sol 328–330; data are from the 14.4 keV channel of the Mössbauer spectrometer. Also shown are (in red) a typical bright soil spectrum from Opportunity (MER B; sol 60) and (in blue) a spectrum of the Compositional Calibration Target (CCT) on Opportunity. These spectra have been scaled to match the intensity of the capture magnet spectrum. The (thin) error bars on each data point in the Mössbauer spectrum show $\pm\sqrt{N}$, where N is the number of counts in each data point.

an increase in peak areas of Fe, Ti and Cr accompanied by a decrease of the area of all other peaks (except Al, which is the material on which the dust is residing).

The Viking biology experiments showed that the martian soil was not only oxidized, but also strongly oxidizing when brought into contact with liquid water¹⁹. This led some to believe that the magnetic mineral must also be highly oxidized, and the mineral maghemite was therefore proposed as the most probable candidate for the cause of the magnetic properties of the dust^{9,20,21}. However, the present results show that the magnetic mineral is less oxidized than hitherto believed.

The minerals in the dust on the magnets have about $45 \pm 10\%$ of the Fe in the ferric state, which is higher than for average martian soil¹⁴. The spectrum contains all the components found in Mössbauer spectra of different soils in both Gusev crater and on Meridiani Planum¹⁸. In this sense, the mineralogical composition of the dust on the magnets does not differ substantially from average martian soil.

A comparison of the Mössbauer spectra of dust on the filter magnet (not shown) with that of the capture magnet shows that a significant amount of a magnetic mineral, magnetite, is present in dust particles on both magnets. Furthermore, both magnets hold significant amounts of particles containing paramagnetic minerals. Mössbauer spectroscopy gives no information about the existence of iron-free particles, such as feldspar. However, within the resolution of the Pancam images the centre of the sweep magnet is clean of dust¹⁰ (unless the particles here have a reflectance identical to aluminium). This indicates that the airborne dust does not contain more than a few per cent of single-phase non-magnetic particles (for example, feldspar), and therefore suggests that all airborne particles are composite. So the magnets are evidently not just culling a subset of single-phase magnetic particles like, say, magnetite (Fe_3O_4). Instead, the results allow the firm conclusion that most of the airborne particles are composite.

A possible interpretation of the results described above is that the airborne dust consists of two groups of minerals: one of primary minerals and one of secondary (possibly hydrated) minerals. The first group consists of olivine, pyroxene and magnetite (possibly titanomagnetite) crystallites eroded from rocks. The second group consists of more oxidized crystallites of ferric and nanocrystalline oxides of which many could be secondary minerals. Some of these minerals may have formed by interaction with water. Crystallites from both of these groups seem to have been assembled into dust grains—perhaps just bound by electrostatic forces. Products of palagonitization, commonly used as terrestrial Mars dust analogues, correspond to such a mixture of particles²². Each particle contains several metallic elements (such as Na, Mg, Al, Si, K, Ca, Ti and Fe), and each particle contains several ferrous and ferric mineral phases.

How did such particles originate? The fact that the suspended particles contain the mineral olivine— $(\text{Mg,Fe})_2\text{SiO}_4$ —shows that the particles cannot be solely a result of precipitation in liquid water. The presence of olivine in the dust points towards physical alteration rather than chemical weathering. On the other hand, the dust particles cannot be exclusively unaltered 'small basaltic rocks' because of the relatively high abundance of the doublet assigned to nanocrystalline ferric oxides and the possible presence of haematite. Some chemical alteration must have taken place, including oxidation of the Fe(II) in the rocks. The findings are consistent with the following scenario.

The primary minerals of the dust particles formed by physical processes (diurnal temperature cycles, comminution by meteoritic impacts, wind abrasion) from parent basaltic rocks. These processes have been operating since the formation of the planet. Some very slow alteration processes, possibly driven by diurnal condensation of thin films of atmospheric water²³, may have occurred simultaneously. The secondary minerals could also (or in addition) be remnants from an early water-rich period in the history of the planet. In that case

they would not be an integral part of the bulk of the dust particles, but would rather be bound electrostatically to their surface. As long as no information on the morphology of the dust particles is available on a microscopic scale, the time of origin of the secondary minerals remains unknown. (We note that a microscope with a resolution better than $10\text{ }\mu\text{m}$ will be on board the Phoenix lander, Mars Scout Mission 2007; ref. 24.) In any case, the secondary Fe(III) compounds appear to be the dominant chromophore component of the dust. The paramagnetic or superparamagnetic iron oxides (probably haematite and maybe oxyhydroxides), which provide the dust with its characteristic reddish colour, are thus different from the mineral (magnetite) that causes the magnetism of the dust.

METHODS

The Mössbauer spectrum of pure stoichiometric magnetite, $\text{Fe(II)}[\text{Fe(II)-Fe(III)}]\text{O}_4$, consists of two superimposed sextet components: one originates from Fe(III) ions tetrahedrally coordinated in the close-packed oxygen structure of the magnetite. The other sextet, originating from octahedrally coordinated Fe ions (indicated by square brackets), has an intermediate isomer shift, between typical values for Fe(III) and Fe(II). The fact that these two oxidation states of iron give rise to only one sextet is caused by the process of rapid electron exchange between the Fe(II) and the Fe(III) in magnetite. The two sextets in stoichiometric magnetite have area ratios of 2:1 for octahedrally to tetrahedrally coordinated iron, respectively. The two sextets coincide at high velocities (about $+5.0\text{ mm s}^{-1}$ and $+8.5\text{ mm s}^{-1}$), causing the characteristic asymmetric appearance of the Mössbauer spectrum of pure stoichiometric magnetite.

Maghemite ($\gamma\text{-Fe(III)}_2\text{O}_3$), which holds tetrahedrally and octahedrally coordinated iron (both in the ferric oxidation state), also has two sextet components, but they differ very little in isomer shift, and therefore appear as one slightly broadened sextet. To the naked eye, a content of maghemite would appear in the Mössbauer spectrum as an increase of the first of the two sextets described for magnetite. This is practically indistinguishable from a partial oxidation of magnetite into what we refer to as non-stoichiometric magnetite. In this case, the oxidation is expressed by an area ratio between the two sextets in magnetite that differs from the stoichiometric ratio of 2:1. Also, substitution of other elements (primarily at the octahedral site) can cause a change in the apparent ratio between the two sextets of magnetite. This change is accompanied by a broadening of the second sextet (in pure magnetite assigned to octahedrally coordinated iron).

Haematite has slightly larger values for the magnetic hyperfine field and the isomer shift as compared to maghemite. In addition, owing to its non-cubic structure, it has a non-zero quadrupole shift, leading to a small asymmetry in the distance between the pairs of lines 1 and 2 and between 5 and 6. The fit of the Mössbauer spectra of the dust on the magnets is improved by the inclusion of a sextet haematite component, indicating a minor content of haematite in the dust.

Received 23 November 2004; accepted 29 April 2005.

- Knudsen, J. M. *et al.* Mössbauer spectroscopy on the surface of Mars. Why? *Hyperfine Interact.* **68**, 83–94 (1991).
- Squyres, S. W. *et al.* The Athena Mars Rover science investigation. *J. Geophys. Res.* **108**(E12), doi:10.1029/2003JE002121 (2003).
- Madsen, M. B. *et al.* Magnetic properties experiments on the Mars Exploration Rover mission. *J. Geophys. Res.* **108**(E12), 8069, doi:10.1029/2002JE002029 (2003).
- Tomasko, M. G., Dose, L. R., Lemmon, M., Smith, P. H. & Wegryn, E. Properties of dust in the Martian atmosphere from the Imager on Mars Pathfinder. *J. Geophys. Res.* **104**, 8987–9007 (1999).
- Lemmon, M. *et al.* Atmospheric imaging results from the Mars Exploration Rovers. *Science* **306**, 1753–1756 (2004).
- Pollack, J. B. *et al.* Properties and effects of dust particles suspended in the Martian atmosphere. *J. Geophys. Res.* **84**, 2924–2945 (1979).
- Squyres, S. W. *et al.* The Spirit Rover's Athena science investigation at Gusev Crater, Mars. *Science* **305**, 794–799 (2004).
- Squyres, S. W. *et al.* The Opportunity Rover's Athena science investigation at Meridiani Planum, Mars. *Science* **306**, 1698–1703 (2004).
- Hviid, S. F. *et al.* Magnetic properties experiments on the Mars Pathfinder Lander: Preliminary results. *Science* **278**, 1768–1770 (1997).
- Bertelsen, P. *et al.* Magnetic properties experiments on the Mars Exploration Rover Spirit at Gusev Crater. *Science* **305**, 827–829 (2004).
- Kinch, K. M. *et al.* Preliminary analysis of the MER magnetic properties experiment using a CFD model. *Planet. Space Sci.* (submitted).
- Gellert, R. *et al.* Chemistry of rocks and soils in Gusev Crater from the Alpha Particle X-ray Spectrometer. *Science* **305**, 829–832 (2004).

13. Rieder, R. *et al.* Chemical composition of Martian rocks and soils at Meridiani Planum from the Alpha Particle X-ray Spectrometer. *Science* **306**, 1746–1749 (2004).
14. Yen, A. *et al.* An integrated view of the chemistry and mineralogy of Martian soils. *Nature* doi:10.1038/nature03637 (this issue).
15. Klingelhöfer, G. *et al.* Athena MIMOS II Mössbauer spectrometer investigation. *J. Geophys. Res.* **108**, 8067–8084 (2003).
16. Morris, R. V. *et al.* Mineralogy at Gusev Crater from the Mössbauer Spectrometer on the Spirit Rover. *Science* **305**, 833–836 (2004).
17. Schröder, C., Klingelhöfer, G. & Tremel, W. Weathering of Fe-bearing minerals under Martian conditions, investigated by Mössbauer spectroscopy. *Planet. Space Sci.* **52**, 997–1010 (2004).
18. Klingelhöfer, G. *et al.* Jarosite and hematite at Meridiani Planum from Opportunity's Mössbauer spectrometer. *Science* **306**, 1740–1745 (2004).
19. Levin, G. V. & Straat, P. A. Viking labeled release biology experiment: Interim results. *Science* **194**, 1322–1329 (1976).
20. Hargraves, R. B., Collinson, D. W., Arvidson, R. E. & Cates, P. M. The Viking magnetic properties experiment: Extended mission results. *J. Geophys. Res.* **84**, 8379–8384 (1979).
21. Madsen, M. B. *et al.* The magnetic properties experiments on Mars Pathfinder. *J. Geophys. Res.* **104**, 8761–8779 (1999).
22. Morris, R. V. *et al.* Phyllosilicate-poor palagonitic dust from Mauna Kea Volcano (Hawaii): A mineralogical analogue for magnetic martian dust? *J. Geophys. Res.* **106**, 5057–5083 (2001).
23. Yen, A. *et al.* Subsurface weathering of rocks and soils at Gusev Crater. *Lunar Planet. Sci. Conf. XXXVI*, 1571 (2005).
24. Smith, P. The Phoenix Mission to Mars. *Lunar Planet. Sci. Conf. XXXV*, 2050 (2004).

Acknowledgements This paper is dedicated to the memory of Jens Martin Knudsen (1930–2005), who inspired and promoted our work with Mars over almost two decades. During his last year, a long-standing aspiration of his came true: a Mössbauer spectrum of martian dust was obtained. The Danish authors wish to thank the Athena Science Team and JPL engineers for supporting and carrying out the sometimes tricky, currently ongoing magnet observations on both rovers. Support from the Danish Research Agency, the Thomas B. Thirge foundation and DELTA Danish Electronics, Light and Acoustics is acknowledged.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to W.G. (goetz@mps.mpg.de)

LETTERS MARS

Water alteration of rocks and soils on Mars at the Spirit rover site in Gusev crater

Larry A. Haskin^{1,‡}, Alian Wang¹, Bradley L. Jolliff¹, Harry Y. McSween², Benton C. Clark³, David J. Des Marais⁴, Scott M. McLennan⁵, Nicholas J. Tosca⁵, Joel A. Hurowitz⁵, Jack D. Farmer⁶, Albert Yen⁷, Steve W. Squyres⁸, Raymond E. Arvidson¹, Göstar Klingelhöfer⁹, Christian Schröder⁹, Paulo A. de Souza Jr¹⁰, Douglas W. Ming¹¹, Ralf Gellert¹², Jutta Zipfel¹², Johannes Brückner¹², James F. Bell III⁸, Kenneth Herkenhoff¹³, Phil R. Christensen⁶, Steve Ruff⁶, Diana Blaney⁷, Steven Gorevan¹⁴, Nathalie A. Cabrol¹⁵, Larry Crumpler¹⁶, John Grant¹⁷ & Lawrence Soderblom¹³

Gusev crater was selected as the landing site for the Spirit rover because of the possibility that it once held a lake. Thus one of the rover's tasks was to search for evidence of lake sediments¹. However, the plains at the landing site were found to be covered by a regolith composed of olivine-rich basaltic rock and wind-blown 'global' dust². The analyses of three rock interiors exposed by the rock abrasion tool showed that they are similar to one another, consistent with having originated from a common lava flow^{3–8}. Here we report the investigation of soils, rock coatings and rock interiors by the Spirit rover from sol (martian day) 1 to sol 156, from its landing site to the base of the Columbia hills. The physical and chemical characteristics of the materials analysed provide evidence for limited but unequivocal interaction between water and the volcanic rocks of the Gusev plains. This evidence includes the softness of rock interiors that contain anomalously high concentrations of sulphur, chlorine and bromine relative to terrestrial basalts and martian meteorites⁹; sulphur, chlorine and ferric iron enrichments in multilayer coatings on the light-toned rock Mazatzal; high bromine concentration in filled vugs and veins within the plains basalts; positive correlations between magnesium, sulphur and other salt components in trench soils; and decoupling of sulphur, chlorine and bromine concentrations in trench soils compared to Gusev surface soils, indicating chemical mobility and separation.

In addition to basaltic rocks and global dust, volcanic gases containing sulphur, chlorine, bromine and other volatiles were likely reactants for materials now covering the Gusev plains. Atmospheric oxidants are solar photoproducts¹⁰ from the action of ultraviolet radiation on CO₂ and H₂O, but volcanogenic SO₃ and HCl provide acidic molecules, which, with the help of liquid water, are capable of dissolving basaltic materials¹¹. Acidic alteration of volcanic rocks in an aqueous environment would result in an increase of oxidation state and changes in concentrations of soluble cations (K, Na, Mg, Ca, Fe²⁺) relative to less-soluble cations (Ti, Si, Al, Cr, Fe³⁺). Sources of small quantities of water include precipitation and condensation from the atmosphere (for example, frost), and 'snow' or ice that might have covered the surface at Gusev¹². Ice might have been

trapped at shallow depths during periods of high obliquity, then melted and risen (in response to warming of the surface) as transient water: at times when the atmospheric water-vapour pressure was high enough, liquid water could have been formed¹³.

The survival of basaltic minerals such as olivine, plagioclase and magnetite in the soils and rock coatings⁶ suggests that chemical reactions took place with a low water/rock ratio. From sol 1 to sol 156, no examples of large-scale conversion of basaltic materials to alteration products were found. Physical processes (for example, impacts and aeolian weathering) are responsible for comminution of the basaltic component of the soil¹⁴. Volatile elements (S, Cl, Br), however, are found inside plains basalts in higher concentrations than in terrestrial basalts^{5,15}. The specific grinding energy used for the hardest plains basalt, Humphrey, is only ~50% of the energy used for typical terrestrial basalts¹⁶.

The coatings and interior features of the light-toned rock Mazatzal (Fig. 1; see Methods) indicate post-crystallization aqueous alteration. The inverse correlation of Fe³⁺ in nanophase ferric oxides (Fe_{np-ox}³⁺) with Fe²⁺ in olivine (Fe_{olivine}²⁺) (Fig. 2) and a positive correlation with SO₃ (ref. 17) are qualitatively consistent with the hypothesis that the source of increased Fe³⁺ was mainly oxidation of Fe_{olivine}²⁺. (Here Fe_{np-ox}³⁺ refers to a group of fine-grained (<10 nm) poorly crystalline phases considered as general alteration products of plains basalts. These include, for example, the superparamagnetic forms of Fe-oxides, oxyhydroxides, sulphates, and the Fe³⁺ pigment in palagonitic tephra^{6,18}.) Furthermore, the coatings on Mazatzal have 2–5 times higher S and Cl concentrations than the rock interior. These coatings also contain the only occurrence of crystalline haematite detected to date on the basaltic plains of Gusev⁶.

The coatings on Mazatzal are not pure evaporates, nor are they common soil, either local or the proposed global soil¹⁹. Their properties require alteration of the silicate component of adhering soil or of the rock itself (especially olivine), the oxidation of Fe²⁺, and the incorporation of S and Cl. Given their mode of occurrence²⁰, light-toned rocks such as Mazatzal were probably buried in the soil (or dirty snow during the periods of high obliquity) when their coatings developed. Liquid water, even if present in small quantity as

¹Department of Earth and Planetary Sciences, Washington University in Saint Louis, Missouri 63130, USA. ²Department of Earth & Planetary Sciences, University of Tennessee, Knoxville, Tennessee 37996, USA. ³Lockheed Martin Space Systems, Littleton, Colorado 80125, USA. ⁴Exobiology Branch, MS 239-4, NASA Ames Research Center, Moffett Field, California 94035, USA. ⁵Department of Geosciences, State University of New York, Stony Brook, New York 11794-2100, USA. ⁶Department of Geological Sciences, Arizona State University, Box 876305, Tempe, Arizona 85287, USA. ⁷JPL, 4800 Oak Grove Drive, Pasadena, California 91109, USA. ⁸Cornell University, 428 Space Science Buildings, Ithaca, New York 14853, USA. ⁹Institut für Anorganische und Analytische Chemie, Johannes Gutenberg-Universität, Staudinger Weg 9, D-55128 Mainz, Germany. ¹⁰Companhia Vale do Rio Doce, 20030-900 Rio de Janeiro, RJ, Brazil. ¹¹NASA JSC, MC KR, 2101 NASA Parkway, Houston, Texas 77058, USA. ¹²Abteilung Kosmochemie, Max-Planck-Institut für Chemie, Postfach 3060, Mainz, Germany. ¹³USGS Flagstaff, 2255 North Gemini Drive, Flagstaff, Arizona 86001, USA. ¹⁴Honeybee Robotics, New York, New York 10012, USA. ¹⁵SETI Institute and Space Science Division, MS 245-3, NASA Ames Research Center, Moffett Field, California 94035, USA. ¹⁶New Mexico Museum of Natural History and Science, 1801 Mountain Road, Albuquerque, New Mexico 87104, USA. ¹⁷Smithsonian Institution, Washington DC 20024, USA.

[‡]Deceased.

an 'acid fog', is needed to concentrate S-correlated elements from the surrounding soils and to transport resulting acidic brine into rocks along fractures or interconnected vugs. The brines reacted with the igneous minerals of the rock surface and any adhering soil to produce the different visible coatings (some are light-toned and some are dark), which are rich in S, Cl and haematite. Multilayer coatings covering ventifact grooves¹⁴ indicate episodic burial and exhumation. Ground water might have been the source of liquid water, or alteration could have occurred during periods of high obliquity¹² from contact of the rock surface with transient surficial liquid water¹³ from snow melt.

The interiors of all three samples of Gusev plain basalts contain partly filled vugs and veins (plates 9c, 12, 13 in ref. 14). On Earth, such features have distinct chemical compositions resulting from interaction with aqueous fluids during low-temperature or hydrothermal alteration, transport, and deposition of soluble salts and zeolites²¹. The interiors of both Humphrey and Mazatzal contain higher concentrations of Br than their exteriors, with the highest Br/Cl ratio found in abraded Mazatzal (Fig. 2). The Pancam visible–near-infrared spectra of fill materials in vugs and veins indicate that they are different from the rock surface dust or local soil³.

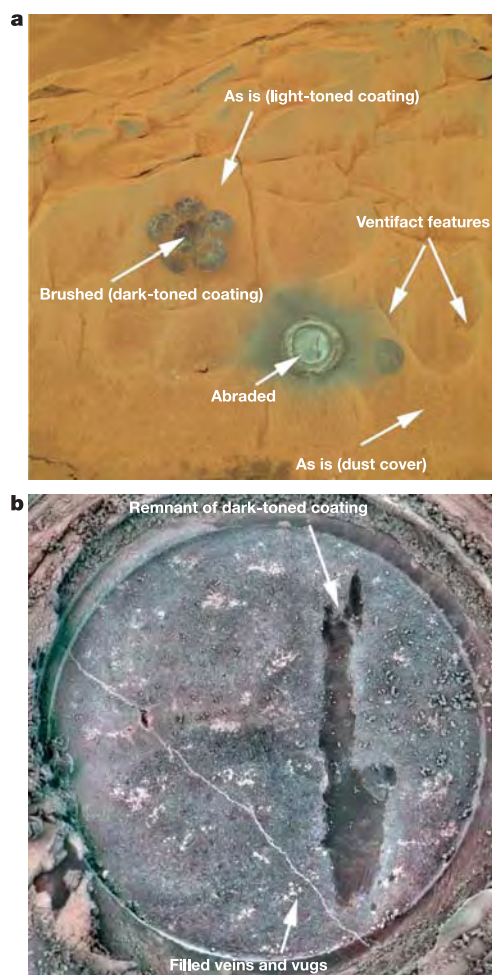


Figure 1 | Rock Mazatzal and its coatings. **a**, Pancam image (P2599 sequence; 480, 530, 600 nm filters) with a 'daisy' brush pattern and an abraded depression. A light-toned yellowish coating occurs beneath the reddish dust cover and atop ventifact grooves and scallops. **b**, Mosaic of four images taken by the Microscopic Imager (merged with Pancam colour data), ~45 mm across. A circular subsurface exposure was ground by the second abrasion. The dark strip (centre right, ~10% area) is a remnant of the dark-toned coating. The basaltic interior shows vugs and veins filled with light-toned materials.

Introduction from volcanic exhalations and the decoupling of Br from Cl owing to the higher solubility of Br in aqueous solution²² are plausible mechanisms to produce the characteristics of alteration observed in these rocks. A network of veins and vugs occurs in the abraded surface of Humphrey (plate 12 in ref. 14). The high salinity (especially Cl) depresses the freezing points of brines²², which would permit penetration into rock interiors at low temperatures. Such brines may become more concentrated by loss of H₂O during their diffusion into the veins and vugs, either by evaporation, addition of soluble compounds, or through chemical reaction with host rocks (especially olivine or glass) to produce hydrated salts, and possibly oxyhydroxides, or hydrated silicates^{23,24}. Sulphates or chlorides would precipitate early, whereas Br-enriched brine would be the last to crystallize; bromides would thus be deposited at the deepest locations reached by the brine. In this mechanism, elevated Br concentrations and higher Br/Cl values in rock interiors would represent the last stages of brines. Compared with the ≤50 p.p.m. Br in typical Gusev soils, the 183 p.p.m. Br found in the interior of Mazatzal would correspond to >1,800 p.p.m. Br in veins and vugs, which constitute 5–10% of the area measured by the APXS (Alpha Particle X-ray Spectrometer).

Soil crusts present beneath the uppermost surface dust¹² suggest some cementing of soil materials by salts deposited from water after the soils were emplaced. Martian humidity and cold night-time temperatures can, in the right seasons and obliquities, produce saturation and even frost or aerosol H₂O. Over geologic timescales, this water could plausibly explain the observed clodding. Evaporating liquid water would also preferentially transport S and Cl towards the surface, explaining the observation that the salts are bound to the soil as precipitates bridging particles, rather than merely mixed with them.

The Laguna trench was dug into the continuous ejecta deposit of Bonneville crater. Imaging and compositional data indicate that the trench site was filled with relatively young basaltic sand deposited by aeolian processes^{3,4,12}. In contrast, the Big Hole and The Boroughs trenches were selected to represent intercrater plains, and were located in topographic lows, in areas with low thermal inertia, away from fresh impact craters and dust-filled hollows. The selection

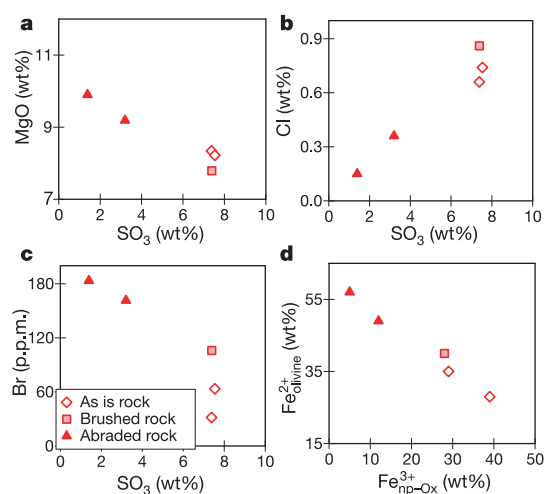


Figure 2 | Chemistry and mineralogy of Mazatzal and coatings. One 'as is' surface target was densely covered by dust and the other was less dusty (exposing the light-toned coating). The brushed surface was entirely covered by the dark-toned coating. The subsurface after the first abrasion was half covered by the dark coating, and the other, ~10% coated. Concentrations of Mg (**a**), Cl (**b**) and Br (**c**) vary with S and span over the range of variation seen in soils (Fig. 3). **d**, The wt% Fe in olivine (as Fe²⁺; Fe²⁺_{olivine}) varies inversely with Fe³⁺ in nanophase ferric oxides (Fe³⁺_{np-Ox}).

was done purposely to seek mature regolith, which may hold a record of more ancient aqueous interactions at Gusev than those expected under current martian surface conditions.

Ratios of four elements (S, Cl, Br and Mg) that are roughly constant in most Gusev soils²⁵ show extended ranges for the soils at both trenches (Fig. 3). The highest concentrations of Mg, S and Br among all Gusev soils occur in the upper-layer soils at The Boroughs trench, accompanied by the highest $\text{Fe}^{3+}/\text{Fe}_{\text{total}}$ value. These characteristics indicate coupling and decoupling of water-soluble compounds of these elements during chemical separations and perhaps transport. Some elemental correlations for the trench soils are strong, even though each trench has only three measurement points. At The Boroughs, a positive correlation exists for SO_3 with MgO, Cl and Br ($r \approx 0.996$, 0.999 and 0.998 ; statistically, $r = 0.988$ is significant at the 90% level for three points); a negative correlation exists for SO_3 with SiO_2 and Al_2O_3 ($r \approx 0.999$ and $r \approx 0.995$); and no obvious correlation exists for SO_3 with CaO and $\text{Fe}_2\text{O}_3(\text{total})$. These characteristics suggest mutual transportation of Mg with the volatile elements (Fig. 3a), and a dilution of basaltic minerals by deposition of salts. These same trends were found in soils at the Viking sites²⁶.

In the trench soils, the concentration of Al_2O_3 decreases sharply and linearly with increasing SO_3 along the same line as for basaltic rocks and their coatings, and the other soils (Fig. 3b). This inverse correlation suggests mixing with an evaporate component having a

high S content. Extrapolation along this trend to zero Al_2O_3 , which appears unaffected by processes other than dilution, yields an SO_3 concentration of ~ 45 wt%. The extrapolated values for all elements are consistent with a mineral assemblage that includes Mg, Ca and Fe sulphates, silica, Ti- and Fe^{3+} -oxides, alkali halides, and phosphate. Moreover, the molar proportions of MgO plus CaO are roughly equal to that of SO_3 , consistent with 5 parts MgSO_4 and 1 part CaSO_4 . Mixing-model calculations suggest ~ 7 – 22 wt% of sulphates (Mg-, Ca- and Fe-sulphate) in The Boroughs soils, and the potential H_2O in hydrated sulphates would constitute up to 4 wt% of the soil in the trench wall (which has the highest Mg and S content).

The SO_3 –Cl and SO_3 –Br correlations for The Boroughs trench soils have significantly different slopes (Fig. 3c, d) from the Big Hole trench soils and common Gusev soils, indicating a decoupling of the three soluble elements. Migration of brine upward, driven by the thermal gradient near the surface, could leave high salt concentrations there. The freezing-point depression of the brine would enable Cl to be carried farther and thus to become decoupled from S; elevated and highly variable Br concentrations argue for episodic migration and evaporation of brines. As an alternative to multiple episodes of brine migration, deposition of impact ejecta could play a role²⁷ in the formation of these compositionally variable soil deposits—highly concentrated salts and strongly altered basaltic regolith (with high $\text{Fe}^{3+}/\text{Fe}_{\text{total}}$) formed at some distant location(s) could have been brought by impact ejecta to become part of the mixture that is now The Boroughs subsurface regolith. In these highly altered basaltic materials, olivine might have been totally consumed and leaching of Ca from clinopyroxene and plagioclase might also have occurred. It is possible that these two mechanisms both contributed to the formation of the observed subsurface regolith at the two trench sites.

Taken together, the data indicate that the geochemical effects of aqueous alteration, though ubiquitous with respect to plains basalts, occurred at low water/rock ratios. Effects on rocks include multiple, oxidized coatings and filled vugs and veins. Effects of aqueous activity are also seen in subsurface soils, including deposition of salts and oxidation of Fe^{2+} . The pattern of alteration of igneous minerals and the deposition and transport of soluble materials implicates interaction with water, but not pools of surface or ground water, or hydrothermal conditions. Small quantities of transient, possibly acidic water could be produced by precipitation and condensation from the atmosphere or by melting of ground ice formed during episodes of high obliquity¹². Water may be involved in the production of the global dust component of the soils, and even the tiny amounts currently present in the martian atmosphere might suffice over time to produce the oxidized material in the dust, with relatively high S and Cl concentrations. More water than at present available from the atmosphere seems to be required, however, to dissolve and mobilize the soluble S-rich components that contributed to the coatings and the salts in subsurface soils. Conditions wetter than at present (and at higher water vapour pressures) are thus implied.

METHODS

Rock and soil targets. Data from three rocks and the soils in three trenches form the basis of discussion in this Letter. Rocks include Adirondack (sols 14–36), Humphrey (sols 54–60) and Mazatzal (sols 76–87). Trenches include Laguna (203 m from Bonneville rim on continuous ejecta deposit, 6–7 cm deep, sols 46–50), Big Hole (556 m from Bonneville, ~ 9 cm deep, sols 113–115) and The Boroughs (1,698 m from Bonneville, ~ 11 cm deep, sols 135–142).

Method of investigation. A full set of analyses¹ was obtained for rock and soil targets discussed in this Letter. These include Pancam multispectral images and Mini-TES spectra; brushing and grinding using the Rock Abrasion Tool (RAT)²⁸; trenching using rover wheels; and microscopic imaging, Mössbauer spectra, and α -particle X-ray spectra for surface analysis before and after the RAT and trench operations.

Rocks and coatings. All rocks investigated from sol 1 to sol 156 at the Spirit site are fine-grained, olivine-rich basalt^{4–8}. Most rocks are angular blocks emplaced

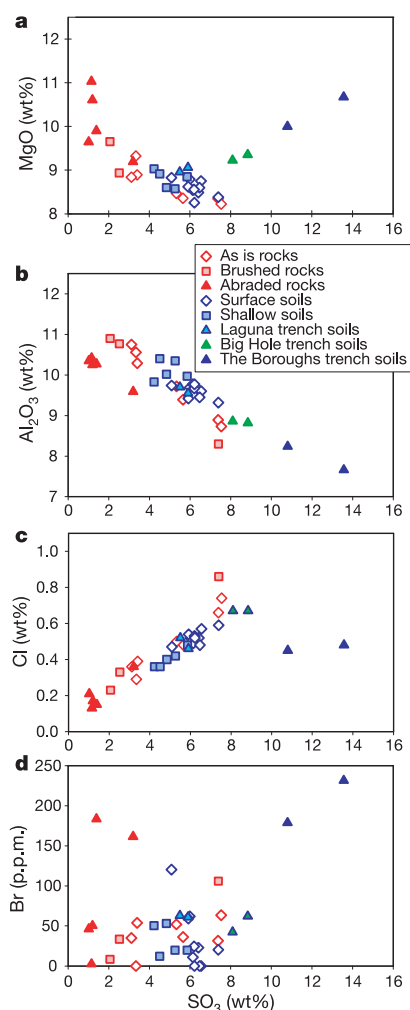


Figure 3 | Concentrations of Mg, Al, Cl and Br versus S in rocks and soils. a–d, Diamond, as is; square, shallow subsurface; triangle, deeper subsurface; red, rocks, blue and green, soils.

as impact ejecta²⁰. Dust cover and surface coatings cause variations in rock brightness in Pancam images³ ranging from dark-toned (for example, Adirondack and Humphrey) to light-toned (for example, Mazatzal)⁹. Dark-toned rocks tend to lie on the surface or are perched, and light-toned rocks tend to be partially buried with low relief²⁰.

Adirondack and Humphrey are relatively coherent and hard, as estimated from the specific grinding energy (SGE) of the RAT¹⁶, although they are softer than terrestrial basalt used for laboratory testing. Microscopic images of rock interiors show dark phenocrysts (interpreted as olivine⁸) within a fine-grained basaltic matrix⁴. Rock interiors contain vugs and veins filled with light-toned materials, distinct from the dust coverings as shown in their Pancam visible–near-infrared spectra³.

The light-toned rock Mazatzal has a similar SGE value to Adirondack¹⁶; both are softer than Humphrey. Mazatzal has a complex multilayer surface coating (Fig. 1). Materials encountered as brushing and abrasion proceeded were: (1) loose dust cover, (2) an outer light-toned soft coating (removed by brushing), (3) a dark-toned hard coating (partially removed by abrasion), (4) an inner light-toned coating, and (5) the basaltic interior matrix with filled vugs and veins. The Mazatzal surface is shaped by ventifact grooves¹⁴, indicating exposure to wind before application of the coatings. A vein cuts across all the coatings except the dust-cover layer.

The dust cover and the coatings on all three rocks have higher concentrations of S and Cl and higher $\text{Fe}^{3+}/\text{Fe}_{\text{total}}$ than the interiors^{5,6} (Figs 2 and 3). The interiors of three rocks have almost identical compositions (except volatiles) and Fe-mineralogy, that is, olivine, pyroxene (or basaltic glass), nanophase ferric oxides ($\text{Fe}_{\text{np-Ox}}^{3+}$), and non-stoichiometric magnetite (ns-Mt), only Mazatzal has a higher $\text{Fe}_{\text{olivine}}^{2+}$ and a lower ns-Mt than others. Concentrations of S in rock interiors (>1 wt%) are higher than is common for basalt (<0.2 wt%)^{15,24}. The dust cover and underlying light-toned and dark-toned coatings on Mazatzal contain crystalline haematite ($\alpha\text{-Fe}_2\text{O}_3$) and the highest $\text{Fe}_{\text{np-Ox}}^{3+}/\text{Fe}_{\text{total}}$ (0.39) analysed by the Spirit Mössbauer spectrometer in rocks thus far⁵ (Fig. 2d). The light-toned and dark-toned coatings of Mazatzal have 2–3 times more S than the surfaces of Adirondack and Humphrey, and the highest concentration of Cl among all rocks and soils examined (Fig. 3c). The interior of Mazatzal has a higher Br concentration than the coatings (Fig. 2c), and much lower $\text{Fe}_{\text{np-Ox}}^{3+}/\text{Fe}_{\text{total}}$ (0.1) without haematite, indicating no residual surface dust from grinding or aeolian infiltration in vein or vug fills.

Soils and trenches. The bulk material of the surface soils is fine-grained, but poorly sorted sand, granules, pebbles and cobbles were observed in the sub-surface soils within some trenches. Slightly cohesive crust was observed beneath the surface dust throughout the Spirit site¹².

Overall, the soils have higher concentrations of S, Cl, P, K and Ti than the interiors of the rocks, and lower Mg, Ca, Cr and Fe (ref. 25). Subsurface soils in rover wheel tracks have lower S and Cl and higher Al (Fig. 3c, b) than the surface dust. There is a site-wide, thin, oxidized upper soil layer (≤ 1 mm thick)^{12,25}, having a slightly higher $\text{Fe}^{3+}/\text{Fe}_{\text{total}}$ value (0.29–0.40) than found in the disturbed soils (0.26–0.27) in the rover wheel tracks. The types of Fe-minerals found in the surface and subsurface soils are the same as those in the plains basalts—that is, olivine, pyroxene (or basaltic glass), $\text{Fe}_{\text{np-Ox}}^{3+}$, and ns-Mt; no haematite was detected⁶.

The subsurface soils exposed by the Laguna trench have nearly identical compositions to the surface soils (Fig. 3), but are less oxidized ($\text{Fe}^{3+}/\text{Fe}_{\text{total}} = 0.22$ compared to 0.3). In contrast, the concentrations of S, Mg and Br, and the $\text{Fe}^{3+}/\text{Fe}_{\text{total}}$ values (0.33–0.44) from the subsurface soils within the Big Hole and The Boroughs trenches are significantly higher than those at the surface (Fig. 3). The highest $\text{Fe}^{3+}/\text{Fe}_{\text{total}}$ and S, Mg, Br concentrations among all Gusev soils were found on the wall of The Boroughs trench.

Received 18 November 2004; accepted 8 April 2005.

1. Squyres, S. W. *et al.* Athena Mars rover science investigation. *J. Geophys. Res.* **108**(E12), 8062, doi:10.1029/2003JE002121 (2003).
2. Squyres, S. W. *et al.* The Spirit Rover's Athena science investigation at Gusev Crater, Mars. *Science* **305**, 794–799 (2004).
3. Bell, J. F. III *et al.* Pancam multispectral imaging results from the Spirit Rover at Gusev Crater. *Science* **305**, 800–806 (2004).

4. Herkenhoff, K. E. *et al.* Texture of the soils and rocks at Gusev Crater from Spirit's microscopic imager. *Science* **305**, 824–826 (2004).
5. Gellert, R. *et al.* Chemistry of rocks and soils in Gusev Crater from the alpha particle X-ray spectrometer. *Science* **305**, 829–832 (2004).
6. Morris, R. V. *et al.* Mineralogy at Gusev Crater from the Mössbauer spectrometer on the Spirit Rover. *Science* **305**, 833–836 (2004).
7. Christensen, P. R. *et al.* Initial results from the mini-TES experiment in Gusev Crater from the Spirit Rover. *Science* **305**, 837–842 (2004).
8. McSween, H. Y. *et al.* Basaltic rocks analyzed by the Spirit Rover in Gusev Crater. *Science* **305**, 842–845 (2004).
9. McSween, H. Y. & Treiman, A. in *Planetary Materials* (ed. Papke, J.) Ch. 6 (Mineralogical Society of America, Washington DC, 1998).
10. Hunten, D. M. Possible oxidant sources in the atmosphere and surface of Mars. *J. Mol. Evol.* **14**, 71–78 (1979).
11. Settle, M. Formation and deposition of volcanic sulfate aerosols on Mars. *J. Geophys. Res.* **84**, 8343–8354 (1979).
12. Arvidson, R. E. *et al.* Localization and physical properties experiments conducted by Spirit at Gusev Crater. *Science* **305**, 821–824 (2004).
13. Richardson, M. & Mischna, M. Long-term evolution of transient liquid water on Mars. *J. Geophys. Res.* **110**, doi:10.1029/2004JE002367 (2005).
14. Greeley, R. *et al.* Wind-related processes detected by the Spirit Rover at Gusev crater, Mars. *Science* **305**, 810–821 (2004).
15. Govindaraju, K. *Geostandards Newsletter* **18** (Special Issue), App. 1, 01–02 (1994).
16. Bartlett, P. W. *et al.* Summary of Rock Abrasion Tool (RAT) results pertinent to the Mars Exploration Rover science data set. *Lunar Planet. Sci. Conf. XXXVI*, abstr. 2292 (2005).
17. Morris, R. V. *et al.* Abundance and speciation of water and sulfate at Gusev crater and Meridiani planum. *Lunar Planet. Sci. Conf. XXXVI*, abstr. 2239 (2005).
18. Morris, R. V. *et al.* Mineralogy, composition, and alteration of Mars Pathfinder rocks and soils: Evidence from multispectral, elemental, and magnetic data on terrestrial analogue, SNC meteorite, and Pathfinder samples. *J. Geophys. Res.* **105**, 1757–1817 (2000).
19. Clark, B. C. *et al.* Chemical composition of Martian fines. *J. Geophys. Res.* **87**, 10059–10067 (1982).
20. Grant, J. A. *et al.* Surficial deposits at Gusev Crater along Spirit Rover traverses. *Science* **305**, 807–809 (2004).
21. Wang, A. *et al.* Raman spectroscopic characterization of a highly weathered basalt: igneous mineralogy, alteration products, and a microorganism. *J. Geophys. Res.* **104**, 27067–27077 (1999).
22. Lide, D. R. *Handbook of Chemistry and Physics* 82nd edn 4–84 (CRC Press, New York, 2001).
23. Gooding, J. L. & Keil, K. Alteration of glass as a possible source of clay minerals on Mars. *Geophys. Res. Lett.* **5**, 727–730 (1978).
24. Gooding, J. L. Soil mineralogy and chemistry on Mars: Possible clues from salts and clays in SNC meteorites. *Icarus* **99**, 28–41 (1992).
25. Yen, A. *et al.* An integrated view of the chemistry and mineralogy of martian soils. *Nature* doi:10.1038/nature03637 (this issue).
26. Clark, B. C. Geochemical components in Martian soils. *Geochim. Cosmochim. Acta* **57**, 4575–4581 (1993).
27. Haskin, L. A. *et al.* On estimating ejecta deposit thicknesses and proportions of materials from distant basins at lunar highland sites. *Meteorit. Planet. Sci.* **38**, 13–33 (2003).
28. Gorevan, S. P. *et al.* Rock Abrasion Tool: Mars Exploration Rover mission. *J. Geophys. Res.* **108**, 8068, doi:10.1029/2003JE002061 (2003).

Acknowledgements Larry A. Haskin, the lead author of this manuscript, died on 24 March 2005. Larry will be remembered by the Athena Science Team for his dedication since the start of the Athena Science Instrument Payload, for his leadership in developing the Mars Microbeam Raman Spectrometer, and for his contributions during the Mars Exploration Rover mission. We thank the NASA Mars programme for support of our participations in the Mars Exploration Rover mission. We thank the Jet Propulsion Laboratory management and the engineering teams for their handling of the rover, sometimes under trying circumstances. We thank K. Kuebler, L. Keszthelyi, M. Blinder and V. Thomas-Holmes for special assistance.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.W. (alianw@levee.wustl.edu).

Dynamic predictive coding by the retina

Toshihiko Hosoya^{1†}, Stephen A. Baccus^{1†} & Markus Meister¹

Retinal ganglion cells convey the visual image from the eye to the brain. They generally encode local differences in space and changes in time rather than the raw image intensity. This can be seen as a strategy of predictive coding, adapted through evolution to the average image statistics of the natural environment. Yet animals encounter many environments with visual statistics different from the average scene. Here we show that when this happens, the retina adjusts its processing dynamically. The spatio-temporal receptive fields of retinal ganglion cells change after a few seconds in a new environment. The changes are adaptive, in that the new receptive field improves predictive coding under the new image statistics. We show that a network model with plastic synapses can account for the large variety of observed adaptations.

Because the physical world is composed of discrete objects with surfaces of fairly uniform reflectance, image points that are close in space or in time tend to have very similar intensities^{1,2}. This redundancy permits efficient encoding of the visual image³. The receptive field of many ganglion cell types shows spatial antagonism between centre and surround, and a biphasic temporal antagonism^{4–6}. For any given image point, neural circuits in the retina predict the local intensity from values at nearby points in space and preceding points in time, and subtract this predicted value from the actual intensity³. Thus, ganglion cells signal not the raw visual image, but the departures from the predictable structure, under the assumption of spatial and temporal uniformity. This difference signal obtained by predictive coding has a much smaller dynamic range than the raw image, and is therefore more suited for transmission through neural fibres with a limited firing rate^{7–9}.

Although these receptive fields produce an efficient encoding of the average visual scene, animals spend considerable time in environments that differ strongly from the average image statistics. For example, sand or gravel surfaces contain finer spatial variation, such that images are correlated only on short scales; forest or tall grass introduce vertical structure, in which two image points are highly correlated if separated vertically but not horizontally. Self-motion of the observer generates optic flow on the retina, in which the intensity at one point of the image is highly correlated with that at a different point later in time. Under all these conditions, the rules for predictive coding are markedly different from the average. An efficient encoder of visual scenes would adapt its strategy accordingly^{10,11}. If this happens in the retina, one should find that ganglion-cell receptive fields change dynamically depending on the correlation structure of the visual environment, in a way that enhances predictive coding and suppresses the dominant spatio-temporal structure in the stimulus.

To test this proposal, we recorded spike trains from ganglion cells in the retinae of salamanders and rabbits. We manipulated the statistics of the visual scene and tested whether adaptation to a different environment altered the encoding of retinal signals. Furthermore, we asked whether these changes conformed to the notion of dynamic predictive coding.

Adaptation to spatial image correlations

Because centre-surround antagonism has been explained as an evolutionary adaptation to positive image correlations in space, we first tested whether negative correlations would alter these receptive

fields. Environment A was designed as a flickering uniform field (Fig. 1a, b) with perfect positive correlation between all image points, whereas environment B was a flickering checkerboard with perfect negative correlation between two sets of neighbouring tiles. After adaptation to environment A or B, an uncorrelated stimulus P was used to probe the ganglion cells' spatio-temporal receptive field. From the responses to P, we computed the sensitivity of each ganglion cell to stimuli of type A or type B (see Methods). For example, the salamander ganglion cell illustrated in Fig. 1c, d experienced a large change in its receptive field. After adaptation to environment B, the receptive field profile flattened, so the neuron was almost equally sensitive to the two checkerboard regions (Fig. 1d). As a result, the cell became less sensitive to stimuli drawn from the checkerboard environment B by a factor of about 0.57, while becoming more sensitive to stimuli from the uniform environment A by a factor of about 1.4. The reverse changes occurred during adaptation to environment A. Figure 1e plots these sensitivity changes caused by adaptation for a large sample of ganglion cells. For most cells, the sensitivity to the adapting stimulus decreased, whereas that to the novel stimulus increased. In some cases the sensitivity to both stimuli changed in the same direction, while still enhancing the novel stimulus. Overall, most data points lie considerably above the diagonal, which means that the cell increased its selectivity for the novel over the adapting stimulus.

The degree to which the novel stimulus is enhanced is summarized by the adaptation index α (equation (2)); note that $\alpha = 1$ if a neuron does not change its sensitivity or changes it equally to stimuli from the two environments, but $\alpha > 1$ if a neuron preferentially suppresses stimuli from the environment to which it is adapted. Across the population of ganglion cells, about half had an adaptation index significantly greater than 1 (Fig. 1f). These cells exhibited dynamic predictive coding. A differential suppression of the adapting stimulus by factors of $\alpha = 2$ or greater was not uncommon. Such adaptations were observed with checkerboard tile sizes of both 200 μm and 400 μm (Fig. 1e), and earlier work suggests that they occur at least over a range of 140–800 μm (ref. 12).

The gain change was not instantaneous after the switch to a new environment, but occurred gradually (Fig. 1g; see also ref. 12). The adaptation index adjusted with a time constant of several seconds, and there may be even slower components beyond the 10-s range that we measured. Because the immediate light response of a ganglion cell is at least tenfold faster than this (Fig. 1d), one can regard this slow

¹Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [†]Present addresses: RIKEN Brain Science Institute, 2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan (T.H.); Department of Neurobiology, Stanford University, Stanford, California 94305, USA (S.A.B.).

adaptation as a gradual modulation of the rules by which the ganglion cell combines information across space. In addition, it is possible that fast changes in sensitivity¹³ occur immediately after the switch and before our first probe measurement (see Supplementary Information).

Adaptation to oriented stimuli

The two environments in Fig. 1 differ substantially in their spatial frequency content. Retinal interneurons with small or large receptive fields will be driven differently by these two stimuli, leading to the potential that local neurons adapt their sensitivity at various sites in the circuit¹². Thus, we tested whether more subtle spatial correlations are also effective in driving adaptation. In Fig. 2a the two environments consist of flickering bar gratings, oriented horizontally or vertically. These two stimuli were perfectly matched for mean intensity, contrast, and spatial and temporal power spectra; in fact, all image statistics were identical except for the difference in orientation. As in Fig. 1b, we exposed the retina to one or the other environment for several seconds, and then probed each ganglion cell's sensitivity to both horizontal and vertical gratings. Adaptation to the oriented gratings produced changes in the receptive fields of many ganglion cells. Figure 2c illustrates a sample cell whose

receptive field acquired a horizontally elongated shape after exposure to the vertical bars. Across many cells, this adaptation systematically had the effect expected from dynamic predictive coding, namely to enhance sensitivity for the novel orientation relative to the adapting orientation (Fig. 2d, e).

Again, one might suspect that individual cells within the retina are driven differentially by the two stimuli and adapt their sensitivity accordingly. For example, an interneuron might be located at an edge between two bars of the grating such that it is strongly stimulated and fatigued by one orientation but not by the other. We tested this by randomly shifting the grating in each stimulus frame (see Supplementary Information and ref. 12). Under these conditions the spatial correlations in the stimulus are maintained, but no neuron consistently lies on a boundary. Still, many ganglion cells adapted their sensitivity in the direction of predictive coding (Fig. 2e).

We performed the same experiments in rabbit retina. Again, about half of the ganglion cells engaged in dynamic predictive coding (Fig. 2f), with differential gain changes similar to those in the salamander. Thus, dynamic pattern adaptation is a shared aspect of retinal function between amphibians and mammals, animals that differ greatly in ecology and physiology but share the challenge of adjusting to a variable visual environment.

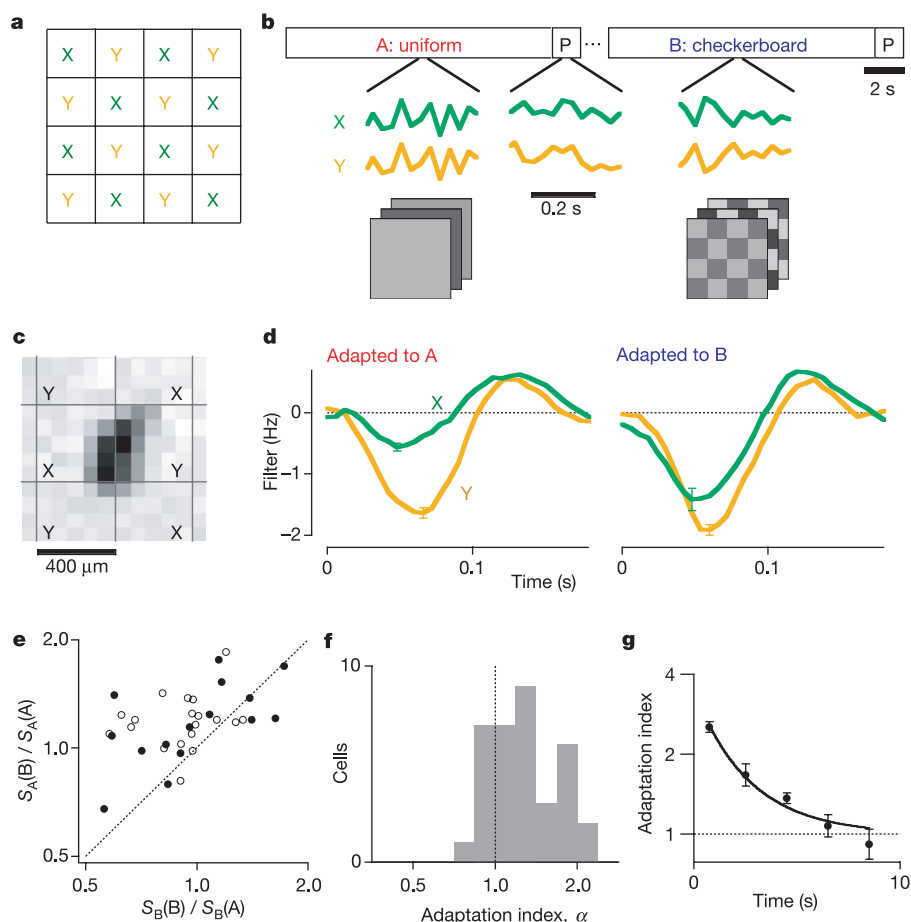


Figure 1 | Adaptation to spatial image correlations. **a**, Stimulus display with square tiles of two intensities X and Y. **b**, Stimulus time course (see Methods). An adapting environment (A or B) was presented for 13.5 s, followed by a probe stimulus (P) for 1.5 s. **c**, Spatial receptive field of a sample ganglion cell, measured by reverse correlation to a random flicker stimulus³⁸. Data in Figs 1, 2, 3 and 4 are from salamander retina, except Fig. 2f, which is from rabbit retina. **d**, Temporal response filter of the ganglion cell to the two stimulus regions X and Y, after adaptation to environment A (left) or B (right). Here and in all other figures, error bars show the s.e.m. of values obtained from independent subsets of the data. **e**, Effects of adaptation to a switch from A to B on the sensitivity S_A for the novel

stimulus A (ordinate) and sensitivity S_B for the adapting stimulus B (abscissa). Each data point represents one ganglion cell from experiments with tile size $400\ \mu\text{m}$ (filled symbols) or $200\ \mu\text{m}$ (open symbols). The dotted line is the identity. The axes are logarithmic. **f**, The adaptation index α (equation (2)) is the change in differential sensitivity resulting from adaptation, and corresponds to the distance of data points above the line in **e**. Histogram of the adaptation index for 35 cells. Of these, 43% had $\alpha > 1$ at $P < 0.05$. **g**, The adaptation index for the ganglion cell shown in **d**, plotted against time in the uncorrelated probe environment, gradually relaxed to 1. An exponential fit to $\log \alpha$ has a time constant τ of 2.7 s. Over seven cells, the time constant was $3.2 \pm 0.8\ \text{s}$ (mean $\pm$ s.e.m.).

Adaptation to temporal and spatio-temporal structure

To test the retina's strategy for predictive coding across time, we constructed two environments matched in intensity and contrast but with very different temporal correlations (Fig. 3a, see Supplementary Information). The intensity at any given time could be predicted by the intensity 60 ms earlier, but the predictive rule had opposite sign in the two environments. In one case (positive correlation) a bright frame was followed 60 ms later by a bright frame, and in the other case (negative correlation) by a dark frame. Adaptation to one environment or the other produced substantial changes in the temporal response function of many ganglion cells. For example, under positive correlations, the cell in Fig. 3b had a strongly biphasic

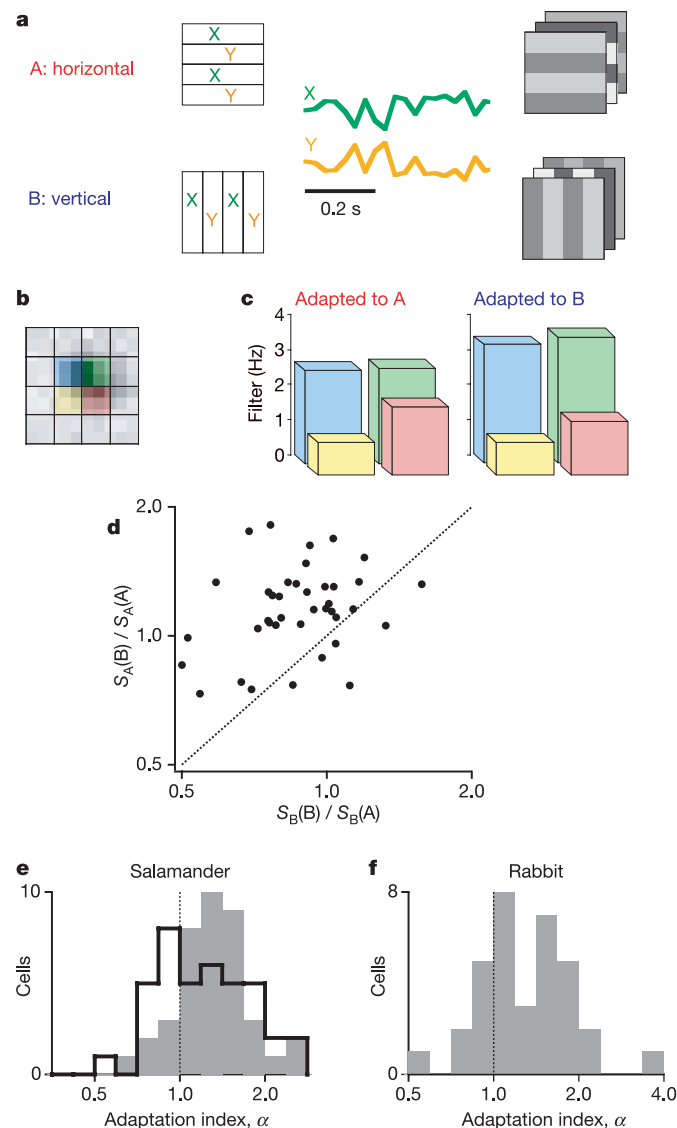


Figure 2 | Adaptation to oriented stimuli. **a**, The environment was a flickering counterphase grating oriented horizontally (A) or vertically (B). The sequence of stimuli was as in Fig. 1b. **b**, Spatial receptive field of a sample ganglion cell. **c**, Sensitivity to the four regions coloured in **b** during the probe interval P, after adaptation to environment A (left) or B (right). For each region the peak amplitude of the response filter is plotted. **d**, Effects of adaptation on the sensitivity to stimuli A and B, displayed as in Fig. 1e. **e**, Histogram of the adaptation index α for 39 cells (grey); 41% of cells had $\alpha > 1$ at $P < 0.05$. In some experiments (line, 40 cells) the borders of the bars were shifted randomly (see Supplementary Information). The means of the two histograms are not significantly different ($P = 0.59$). **f**, Histogram of the adaptation index α for 34 ganglion cells from rabbit retina; 38% of cells had $\alpha > 1$ at $P < 0.05$.

response function, which suppressed its response to stimuli with positive correlation. After adaptation to the other environment, the response function became less biphasic. Many cells underwent similar changes, with the effect being again a suppression of the predictable stimulus features compared with novel features (Fig. 3c, d).

Finally, we tested the retina's ability for predictions across space and time. The stimulus was a flickering checkerboard (Fig. 1a) in which one set of tiles was modulated with a time-shifted version of the flicker in another set of tiles (Fig. 3e). Thus certain points on the retina could serve to predict the intensity at other points, but the order of the time shift, and thus the direction for prediction, was inverted in the two environments. This is a somewhat abstract stimulus, and the difference between the two environments is subtle and almost imperceptible to the human observer. Nevertheless, many retinal ganglion cells changed their response properties significantly from one environment to the other (Fig. 3f, g), leading to a different temporal summation of the intensity from the two stimulus regions. When this occurred, it again had the effect of suppressing predictable features over novel features (Fig. 3h), but the magnitude of the effect was somewhat weaker than for the other stimulus correlations we had tested.

A pattern detector hypothesis

How can the retina accomplish this diverse set of adaptations? A popular hypothesis for phenomena of pattern adaptation postulates that there are several parallel pathways within the circuit that combine to make the output signal^{14,15}, in this case at the retinal ganglion cell (Fig. 4a). Within each pathway, the interneurons are pattern detectors, selective for a particular stimulus feature; for example, bars of a specific orientation. If that feature occurs frequently, those interneurons will be driven strongly, leading to their fatigue and diminished contribution to the output signal. As a result, the output becomes less sensitive to common stimulus features and more sensitive to rare features, for example bars of the orthogonal orientation.

In the retina, the most plausible candidates for parallel interneurons pooled by ganglion cells are the bipolar cells. To explain how ganglion cells adapt to the orientation of a grating (Fig. 2), the bipolar cells would need to be significantly orientation-selective and there should be a range of bipolar cells with different selectivities feeding each ganglion cell. Using intracellular recordings, we measured the receptive fields of ten salamander bipolar cells directly (Fig. 4b) and found that they were round or only slightly elongated (Fig. 4c). Furthermore, the receptive-field centres were small compared with the bars of the gratings, and most of them fell entirely within a single bar. Considering all possible positions and orientations of the receptive field, this panel of bipolar cells was selective for one grating over the other by only a factor 1.06 ± 0.07 (mean $\pm$ s.d.). Moreover, bipolar cells adapt their gain only slightly (about 10%; refs 13, 16), even with a strong change in input amplitude. Altogether, this suggests that adaptation in oriented bipolar cells would contribute less than a 1% orientation selectivity to the ganglion cells, much smaller than the observed gain changes by factors of 1.5–2.5 (Fig. 2d, e).

Significant orientation selectivity has been reported in some other retinal interneurons, notably amacrine cells¹⁷. However, these neurons are primarily inhibitory. In that case, the fatigue of an amacrine cell would lead to less suppression of the adapting stimulus in the ganglion cells, contrary to what we observed. Still, the role of pattern-selective interneurons in these adaptation phenomena deserves further attention.

A network plasticity hypothesis

An alternative explanation invokes plasticity of synapses rather than fatigue of interneurons. Consider a ganglion cell with a centre-surround receptive field (Fig. 5a, left), in which the surround inhibition is conveyed by amacrine cells¹⁸. Suppose that

the inhibitory synapse from an amacrine cell to a ganglion cell is plastic in the following manner: when signals in the presynaptic and postsynaptic neurons are correlated, the synapse becomes stronger; if the two neurons are anticorrelated, the synapse becomes weaker. Note that this is an anti-hebbian rule for plasticity, to be distinguished from the hebbian form, in which correlated activity leads to increased excitation. If one applies a vertical grating stimulus (Fig. 5a, middle), the ganglion cell will be strongly correlated with the amacrine cells above and below, and anticorrelated with those to the left and right. Thus, the synapses will change so as to strengthen inhibition from above and below, and weaken inhibition from the sides. This gives the ganglion cell a receptive field with distinct horizontal orientation, and makes the neuron more sensitive to horizontal gratings and less sensitive to vertical ones.

We formalized this idea as a simple feedforward network from a layer of bipolar cells to ganglion cells (Fig. 5b). In this model, a ganglion cell receives excitatory synapses of fixed strength from bipolar cells, and these determine the default shape of the receptive field in absence of stimulation. In addition, there are inhibitory connections—mediated by amacrine cells—that are plastic as determined by the above correlation-based rule (equations (3) and (4); see Methods). Each ganglion cell integrates its bipolar cell inputs linearly, weighted by synaptic strength. Exposure of the network to different stimulus environments induces change in the amacrine synapses, which in turn alters the receptive field of the ganglion cell. In the approximation of linear processing, one can solve the dynamics of this network analytically (see Methods).

To illustrate the results, we consider a model ganglion cell connected to a 4×4 patch of bipolar cells (Fig. 5c); only the centre 2×2 bipolar cells provide direct excitatory connections, but all bipolar cells make modifiable inhibitory connections through amacrine cells. When the network is exposed to the kinds of stimulus used in the experiments of Figs 1 and 2, the ganglion-cell receptive field changes in a way that suppresses the correlated components of the stimulus. For example, when driven by a spatially uniform stimulus, the receptive field strengthens the antagonistic surround; when driven by a horizontal grating, the receptive field develops a vertical orientation that suppresses sensitivity to horizontal bars; when driven by an uncorrelated stimulus, the receptive field becomes

attenuated uniformly, which is akin to contrast adaptation observed previously^{12,13,19}. More generally, one can show that such an adaptive network performs a multidimensional scaling in the space of stimuli: the sensitivity for any given axis in stimulus space is scaled down according to the strength of stimulation along that axis (see Supplementary Information). Figure 5d illustrates how this stimulus selectivity changes over time in the course of adaptation. The model makes the intriguing prediction that adaptation to an increase in stimulus strength occurs more rapidly than to a decrease, a feature that is indeed consistently observed in animals from flies to humans^{12,20,21}.

Several features distinguish this adaptive network hypothesis (Fig. 5a) from that of adaptive pattern detectors (Fig. 4a). First, adaptation happens at each synapse, not in each neuron. Because there are far more synapses than neurons, this allows a rich set of adaptations. By contrast, the pattern detector model requires a specific interneuron selective for each of the various types of correlation to which the retina can adapt (Figs 1–3). Second, amacrine cells come in a great variety of types, with differing receptive field sizes, integration times, and latencies^{22,23}. This means that many different kinds of stimulus correlation across space and time can be sensed and exploited for predictive coding. Last, this hypothesis predicts that inhibition is essential for the adaptation process, because it implements the subtraction of predictive signals. We tested this by repeating experiments on orientation-adaptation while blocking the inhibitory neurotransmitters GABA (γ -aminobutyric acid) and glycine²³. These conditions profoundly influence retinal signalling: the firing rate of ganglion cells increases and their stimulus selectivity changes as the antagonistic surround weakens^{18,24}. Still, one can ask whether these stimulus selectivities are altered by adaptation, and the adaptation index α (equation (2)) is independent of any absolute changes in sensitivity. We found that, without inhibition, the ganglion cells indeed lost the ability for dynamic adjustment of stimulus selectivity (Supplementary Fig. S3).

Discussion

Although the notion of dynamic predictive coding captures the essence of these effects, we also noted several departures from the

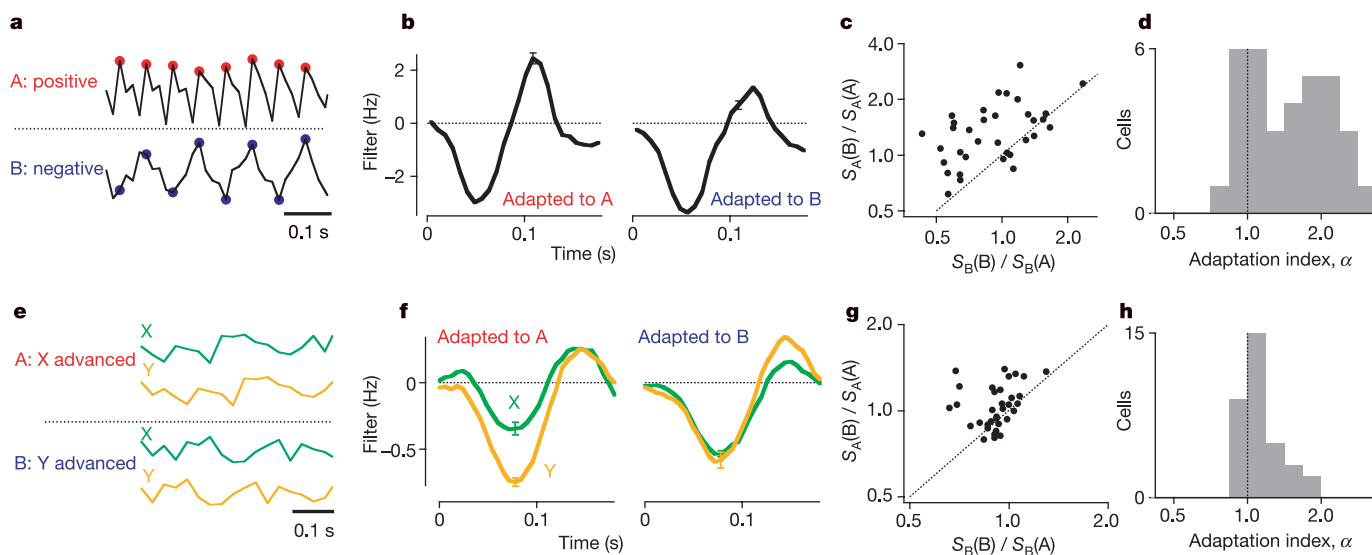


Figure 3 | Adaptation to temporal and spatio-temporal correlations. **a**, The environment was a flickering uniform field whose intensity values 60 ms apart (see markers) had a strong positive (A) or negative (B) correlation. **b**, Filter waveform of a sample ganglion cell, after adaptation to environment A (left) or B (right). **c**, Effects of adaptation on the sensitivity to stimuli A and B, displayed as in Fig. 1e. **d**, Histogram of the adaptation index α for 34 cells; 44% had $\alpha > 1$ at $P < 0.05$. **e**, Environments with

different spatio-temporal correlations. Stimulus display as in Fig. 1a, but with Y the same as X delayed by 60 ms (A) or vice versa (B). **f**, Response kernel of a sample ganglion cell to X and Y, after adaptation to A (left) or B (right). **g**, Effects of adaptation on the sensitivity to stimuli A and B, displayed as in Fig. 1e. **h**, Histogram of the adaptation index α for 34 cells; 41% had $\alpha > 1$ at $P < 0.05$.

theoretically ideal encoder. First, only about half of the observed ganglion cells adapted to each stimulus, and a breakdown by cell type in the salamander retina revealed some systematic differences: in particular, 'fast OFF' cells²⁵ adapted to all the conditions tested, whereas 'slow OFF' cells adapted only to spatial correlations in the stimulus (Figs 1 and 2). Second, the absolute magnitude of adaptation was lower than for the ideal efficient encoder. In theory, a perfectly predictable stimulus component could be completely suppressed, whereas the largest differential gain changes we observed were a factor of 3. Last, some stimulus correlations induced greater adaptation than others (compare Figs 3d with Fig. 3h). All these limitations may provide clues to the circuit mechanisms that underlie the effects. For example, under the adaptive network hypothesis (Fig. 5) the range of conditions to which a given ganglion cell type can adapt is limited by which classes of amacrine cells it contacts in the inner plexiform layer. The degree of adaptation is limited by how sensitive an amacrine cell synapse is to the correlation signal, for example the parameter β in equation (4). Further exploration of the range of adaptive behaviours will certainly be instructive.

With regard to the mechanisms for these adaptive effects, we considered two alternative explanations that place the changes either at individual interneurons (Fig. 4) or at individual synapses (Fig. 5). The former scheme invokes a pattern-selective interneuron for each type of stimulus pattern to be sensed, and fatigue of those interneurons in the course of adaptation. This interpretation is widespread in the literature, to the point at which the observation of pattern adaptation is accepted as evidence for the existence of pattern detector neurons^{14,15}. For example, in human vision, adaptation to an oriented grating raises the contrast threshold for that same stimulus by a factor of 2–4 (refs 26, 27). This is commonly thought to occur in the visual cortex^{14,27–29}, because it is there that one first encounters

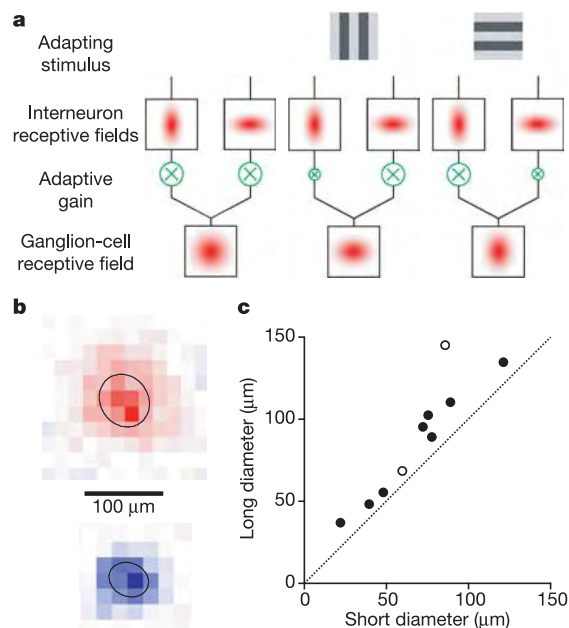


Figure 4 | Pattern detector model for adaptation. **a**, Parallel pattern-selective pathways converge at the ganglion cell and adjust their gain independently. Each pathway contains orientation-selective interneurons (left) whose response becomes fatigued under prolonged exposure to the preferred orientation (middle and right). **b**, Receptive field profiles of two salamander bipolar cells (top, ON type; bottom, OFF type). Each profile was fitted with a gaussian shape, and the line shows the contour of the gaussian at 1 s.d. from the centre. **c**, Receptive field shape of 10 salamander bipolar cells (open symbols, ON type; closed symbols, OFF type), given by the long and short diameters of the 1 s.d. ellipse from the gaussian fit.

overtly orientation-selective cells. However, given that many retinal ganglion cells adapt to oriented gratings with a gain change of 1.5–2.5 (Fig. 2d, e), it is possible that half of the psychophysical after-effect already arises in the retina. Furthermore, given the extensive knowledge of retinal circuits and the lack of orientation selectivity in retinal bipolar cells (Fig. 4b), this is unlikely to occur through fatigue of pattern-selective interneurons. Thus, one may need to reconsider both the site and the mechanism of various perceptual adaptation effects in human vision.

The alternative hypothesis of a modifiable network (Fig. 5) is intriguing, because a single assumption for plasticity at amacrine cell synapses can explain a host of seemingly different adaptations

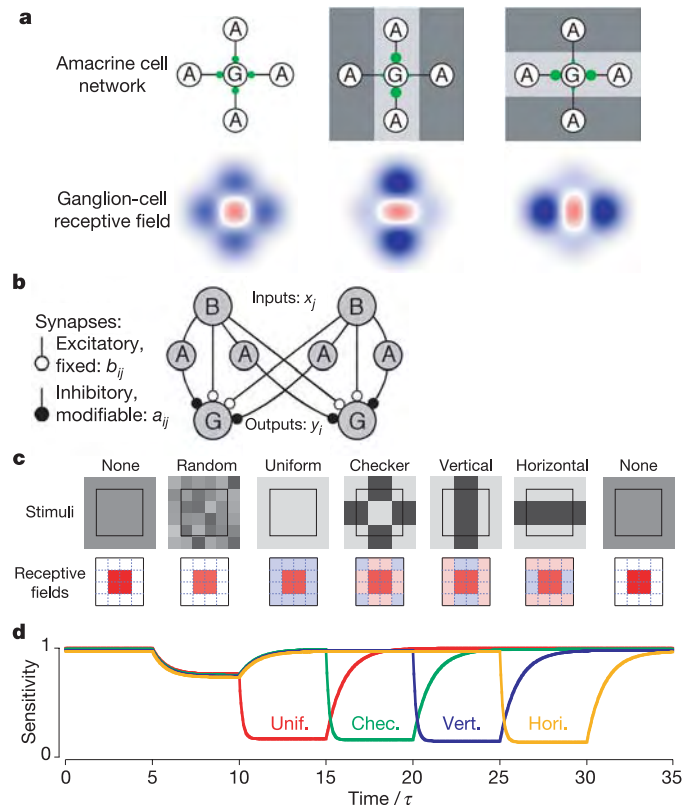


Figure 5 | Network plasticity model for adaptation. **a**, Inhibitory amacrine cells connect to a ganglion cell through modifiable synapses (top), providing the receptive field's antagonistic surround (bottom). Under patterned stimulation (middle and right) each synapse changes strength depending on the correlation between presynaptic and postsynaptic activity. This shapes the receptive field in a way that reduces sensitivity to the predominant stimulus pattern. **b**, Schematic circuit of the inner retina. Bipolar cells (B), carrying input signals x_j , connect to ganglion cells (G) with output signals y_i , through fixed excitatory synapses b_{ij} and through modifiable inhibitory synapses a_{ij} of amacrine cells. **c**, Adaptive change in the receptive field of a model ganglion cell connected to a 4×4 array of bipolar cells. Top, sequence of stimulus environments driving the adaptation: steady grey screen, independently flickering pixels, flickering uniform field (Fig. 1b), flickering checkerboard (Fig. 1b), flickering vertical bars (Fig. 2a), flickering horizontal bars (Fig. 2a), steady grey screen. The square outline marks the 4×4 pixels in the ganglion-cell receptive field; each pixel drives one bipolar cell. Bottom, receptive field profile of the ganglion cell after adaptation to each environment. Colours reflect the net synaptic connectivity to each pixel, $a_{ij} + b_{ij}$; red is positive, blue is negative. The fixed connections b_{ij} are limited to the central four bipolar cells. Pattern adaptation induces contributions from the surrounding pixels in a manner that suppresses the adapting stimulus. **d**, Time course of adaptation. Each environment is applied for a period of five adaptation time constants τ . Curves show the sensitivity of the ganglion cell (see equation (12) in the Supplementary Information) to four different types of stimulus: uniform (unif.), checkerboard (chec.), vertical (vert.) and horizontal (hori.).

(Fig. 5c, d). The qualitative function of this network has a simple interpretation: one can view each amacrine cell as ‘trying to predict’ the response of the ganglion cell from stimulus information in its own spatio-temporal receptive field. Those neurons that are successful get ‘rewarded’ with a stronger inhibitory synapse. As a result, the successful predictions are subtracted from the ganglion cell input, and by slowly adjusting its synapses the network literally performs a dynamical version of predictive coding. A precedent for such an adaptation mechanism is found in the electrosensory system of weakly electric fish, in which the prediction for sensory input is performed using efferent motor signals³⁰. Again, the predictions are adjusted dynamically, in a recurrent network that has been shown to use activity-dependent synapses with anti-hebbian plasticity. Plausibility notwithstanding, the central ingredient of this hypothesis is the modifiable amacrine cell synapse and its dependence on presynaptic/postsynaptic correlation. This kind of use-dependent plasticity has been reported for inhibitory synapses in the cerebellum³¹ and elsewhere³², but it remains to be explored in the retina. Finally, the two mechanisms considered here—changes in neuronal sensitivity versus changes in connectivity—are not exclusive, and may each contribute to adaptation.

Faced with the remarkable plasticity of retinal processing—light adaptation³³, contrast adaptation³⁴ and pattern adaptation—the question arises how the brain copes with the continually changing rules of encoding in the sensory periphery. For the fly eye, it has been suggested that the state of adaptation is communicated downstream through certain slow statistics of visual spike trains²¹. However, there is no necessity to inform the brain of adaptive changes in coding. The goal of the visual system is not to construct internally a veridical reproduction of the intensity pattern on the retina. Instead, the system must reduce the onslaught of raw visual information and extract the few bits of information that are relevant to behaviour. This entails the discarding of signals that are less useful. The many known visual illusions of brightness and contrast are evidence of such information loss, as are the illusory after-effects of pattern adaptation. Thus, pattern adaptation is not merely a scheme for efficient recoding but rather serves to strip from the visual stream predictable and therefore less newsworthy signals.

METHODS

Recording. The dark-adapted retina of a larval tiger salamander or New Zealand white rabbit was isolated under an infrared microscope into oxygenated Ringer’s medium (salamander, 22 °C) or Ames’ solution (rabbit, 37 °C). A piece of retina, 2–4 mm (salamander) or about 5 mm (rabbit) on a side, was placed with the ganglion cell side downwards onto an array of 61 electrodes to record action potentials from ganglion cells, as described previously^{12,35}. For intracellular recordings from salamander bipolar cells¹³, sharp microelectrodes were filled with 2 M potassium acetate and 1% Alexa 488, having a final impedance of 150–250 MΩ. After recording, cells were filled ionophoretically (–1 to –2 nA pulses, about 10–15 min), and imaged with a 40 × water-immersion objective.

Stimulation. Visual stimuli were generated on a computer monitor and projected through an objective lens onto a 3.25-mm diameter aperture of the retina. The light was white, with a mean intensity of $M \approx 4 \text{ mWm}^{-2}$ at the retina, in the regime of photopic vision. All experiments employed random flicker stimuli, whose single-point statistics were identical for every location on the retina. Over time, each point experienced intensity values distributed as a gaussian with mean M and standard deviation $C = 0.35M$. The intensity was updated with another random value at periodic intervals of 15 or 30 ms. The various stimulus environments differed in their two-point statistics, namely the correlation between the intensity values at different points in space or time.

The goal was to probe retinal response properties after adaptation to two different stimulus environments A and B. The probe environment P was always neutral, in that it contained a superposition of stimuli of types A and B. We interleaved segments of the two adapting stimuli and the probe stimulus in the order

$$A_1 P_1 A_2 P_2 \dots A_{10} P_{10} B_1 P_{11} B_2 P_{12} \dots B_{10} P_{20} A_{11} P_{21} A_{12} P_{22} \dots A_{20} P_{30} \dots$$

where A_i , B_i and P_i represent stimulus segments with different pseudo-random flicker sequences drawn from environments A, B and P. In a typical experiment, 500 different segments of P were collected for each adaptation state. The probe

segments were kept short (1.5 s) relative to the intervening adapting segments (13.5 s). For the experiment of Fig. 1g, the probe lasted 10 s and the adaptation 50 s. Over several hours of experimenting, the ganglion cell firing rates remained fairly stable, with a typical variation of 20%. Through the rapid interleaving of different stimuli, the analysis was robust to any slow drifts in response properties.

The Supplementary Information gives details on the construction of the stimuli.

Analysis. The visual responses of retinal ganglion cells can be approximated well by a linear–nonlinear (LN) model³⁶. This is a simple mathematical functional that turns a visual stimulus into the neuron’s firing rate. The stimulus is passed through a spatio-temporal linear filter, and the resulting variable is transformed by a nonlinear function that can account for firing threshold and saturation. For example (Supplementary Fig. S2), if the stimulus contains two spatial regions with intensity time courses $x(t)$ and $y(t)$, the cell’s firing rate $r(t)$ is

$$r(t) = N(g(t)) = N\left(\int x(t')L_X(t-t')dt' + \int y(t')L_Y(t-t')dt'\right) \quad (1)$$

where $L_X(\tau)$ and $L_Y(\tau)$ are the time-dependent impulse responses of the filters applied to stimulus variables x and y respectively, and $N(g)$ is the nonlinearity.

The brief probe segments P were used to derive the best-fit LN model for each of the adapting conditions A and B. From the measured response to gaussian random flicker, we estimated both the filter and the nonlinearity by standard reverse-correlation algorithms³⁶. The waveforms of these filters for representative neurons are plotted in Figs 1d, 3b and 3f.

To assess the degree of predictive coding, we evaluated the sensitivities of the neuron’s spatio-temporal filter to stimuli drawn from environments A and B, denoted S_A and S_B , respectively (see Supplementary Information). The filter computed under adapting condition A yielded sensitivities $S_A(A)$ and $S_B(A)$, and the filter under adapting condition B yielded $S_A(B)$ and $S_B(B)$. In the course of adaptation to B, the sensitivity S_A changes by a factor $S_A(B)/S_A(A)$, and S_B changes by a factor $S_B(B)/S_B(A)$. These factors are plotted in Figs 1e, 2d, 3c and 3g. The ratio of these two factors is the adaptation index

$$\alpha = \frac{S_A(B)/S_A(A)}{S_B(B)/S_B(A)} \quad (2)$$

which is plotted in Figs 1f, 1g, 2e, 2f, 3d and 3h. It measures the extent to which adaptation selectively suppresses the adapting stimulus. The hypothesis of dynamic predictive coding predicts that $\alpha > 1$. For each cell, the experimental uncertainty in α was determined using independent subsets of data, and the P value for $\alpha > 1$ was computed by means of a one-tailed t -test.

Pharmacology. After the addition of 10 μM strychnine and 100 μM picrotoxin to the bathing solution, ganglion cell firing rates doubled on average. The light responses did not saturate, as judged by the form of the nonlinearity in the LN fits. After return to control solution, we did not achieve full reversal of the drug effects within the available time; washout of these drugs from a whole-mount preparation is exceedingly slow³⁷.

Bipolar cell receptive fields. The spatio-temporal receptive field of each bipolar cell was measured by reverse-correlating the membrane potential to a flickering checkerboard stimulus³⁵, and then approximated as the product of a spatial profile and a temporal filter³⁸. The spatial profile was fitted by a two-dimensional gaussian bell, characterized by the long and short axis of the ellipse at 1 s.d. (Fig. 4b). To estimate how selective such a cell would be for oriented gratings, the gaussian was convolved with two grating stimuli: one whose bars were aligned with the long axis of the ellipse, yielding the maximal sensitivity, and the other aligned with the short axis, yielding the lowest sensitivity. The ratio of the two sensitivities was taken as the cell’s orientation selectivity, quoted in the text.

Anti-hebbian retina model. Consider a linear feedforward network as illustrated in Fig. 5b. The input layer represents bipolar cells, the output layer ganglion cells. Each ganglion cell receives two kinds of synaptic input from bipolar cells: excitatory synapses with fixed strength b_{ij} , and inhibitory synapses—through intermediate amacrine cells—with a variable strength a_{ij} . The fixed excitatory synapses set the ‘default’ receptive field of a ganglion cell; the inhibitory synapses can weaken or strengthen depending on the stimulus history, and modify the receptive field accordingly.

We suppose that the network operates linearly. For simplicity, we also ignore the dynamics of the light response, and treat retinal processing as instantaneous. If x_j is the activity of bipolar cell j , and y_i the activity of ganglion cell i , then

$$y_i = \sum_j (b_{ij} + a_{ij})x_j \quad (3)$$

The network undergoes adaptation through slow modulation of the synapses a_{ij} following the rule

$$da_{ij}/dt = (-a_{ij} - \beta(y_i x_j))/\tau \quad \tau, \beta > 0 \quad (4)$$

Here $-a_{ij}/\tau$ is a decay term, to ensure that the network does not remember stimulus history forever. The term $-\beta\langle y_i x_j \rangle/\tau$ is dependent on recent activity, specifically the correlation between bipolar cell j and ganglion cell i . Adaptation is driven by the statistics of the inputs x_j to the network. Thus, a change in the stimulus ensemble leads to a change in the ganglion cell's synapses and thus its receptive field. For the solution of these dynamic equations, and their application in Fig. 5c, d, see Supplementary Information.

Received 5 September 2004; accepted 27 April 2005.

- Dong, D. W. & Atick, J. J. Statistics of natural time-varying images. *Network* **6**, 345–358 (1995).
- Field, D. J. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A* **4**, 2379–2394 (1987).
- Srinivasan, M. V., Laughlin, S. B. & Dubs, A. Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond. B* **216**, 427–459 (1982).
- Kuffler, S. W. Discharge patterns and functional organization of mammalian retina. *J. Neurophysiol.* **16**, 37–68 (1953).
- Barlow, H. B. Summation and inhibition in the frog's retina. *J. Physiol. (Lond.)* **119**, 69–88 (1953).
- Meister, M. & Berry, M. J. II The neural code of the retina. *Neuron* **22**, 435–450 (1999).
- Barlow, H. B. in *Sensory Communication* (ed. Rosenblith, W. A.) 217–234 (MIT Press, Cambridge, Massachusetts, 1961).
- van Hateren, J. H. Real and optimal neural images in early vision. *Nature* **360**, 68–70 (1992).
- Atick, J. J. & Redlich, A. N. What does the retina know about natural scenes? *Neural Comput.* **4**, 196–210 (1992).
- Barlow, H. & Földiák, P. in *The Computing Neuron* (eds Durbin, R., Miall, C. & Mitchison, G.) 54–72 (Addison-Wesley, Wokingham, 1989).
- Barlow, H. B. in *Vision: Coding and Efficiency* (ed. Blakemore, C.) 363–375 (Cambridge Univ. Press, Cambridge, 1990).
- Smirnakis, S. M., Berry, M. J., Warland, D. K., Bialek, W. & Meister, M. Adaptation of retinal processing to image contrast and spatial scale. *Nature* **386**, 69–73 (1997).
- Baccus, S. A. & Meister, M. Fast and slow contrast adaptation in retinal circuitry. *Neuron* **36**, 909–919 (2002).
- Graham, N. V. S. *Visual Pattern Analyzers* (Oxford Univ. Press, New York, 1989).
- Mollon, J. D. in *The Perceptual World* (eds Von Fendit, K. & Monstgaard, I. K.) 71–97 (Academic, London, 1977).
- Rieke, F. Temporal contrast adaptation in salamander bipolar cells. *J. Neurosci.* **21**, 9445–9454 (2001).
- Bloomfield, S. A. Orientation-sensitive amacrine and ganglion cells in the rabbit retina. *J. Neurophysiol.* **71**, 1672–1691 (1994).
- Cook, P. B. & McReynolds, J. S. Lateral inhibition in the inner retina is important for spatial tuning of ganglion cells. *Nature Neurosci.* **1**, 714–719 (1998).
- Chander, D. & Chichilnisky, E. J. Adaptation to temporal contrast in primate and salamander retina. *J. Neurosci.* **21**, 9904–9916 (2001).
- Snippe, H. P., Poot, L. & van Hateren, J. H. Asymmetric dynamics of adaptation after onset and offset of flicker. *J. Vis.* **4**, 1–12 (2004).
- Fairhall, A. L., Lewen, G. D., Bialek, W. & van Steveninck, R. R. D. Efficiency and ambiguity in an adaptive neural code. *Nature* **412**, 787–792 (2001).
- Masland, R. H. The fundamental plan of the retina. *Nature Neurosci.* **4**, 877–886 (2001).
- Yang, C. Y., Lukasiewicz, P., Maguire, G., Werblin, F. S. & Yazulla, S. Amacrine cells in the tiger salamander retina: morphology, physiology, and neurotransmitter identification. *J. Comp. Neurol.* **312**, 19–32 (1991).
- Cook, P. B., Lukasiewicz, P. D. & McReynolds, J. S. Action potentials are required for the lateral transmission of glycinergic transient inhibition in the amphibian retina. *J. Neurosci.* **18**, 2301–2308 (1998).
- Warland, D. K., Reinagel, P. & Meister, M. Decoding visual information from a population of retinal ganglion cells. *J. Neurophysiol.* **78**, 2336–2350 (1997).
- De Valois, K. K. Spatial frequency adaptation can enhance contrast sensitivity. *Vision Res.* **17**, 1057–1065 (1977).
- Blakemore, C. & Campbell, F. W. On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *J. Physiol. (Lond.)* **203**, 237–260 (1969).
- Movshon, J. A. & Lennie, P. Pattern-selective adaptation in visual cortical neurones. *Nature* **278**, 850–852 (1979).
- Snowden, R. J. & Hammett, S. T. Subtractive and divisive adaptation in the human visual system. *Nature* **355**, 248–250 (1992).
- Bell, C. C. Memory-based expectations in electrosensory systems. *Curr. Opin. Neurobiol.* **11**, 481–487 (2001).
- Aizenman, C. D., Huang, E. J., Manis, P. B. & Linden, D. J. Use-dependent changes in synaptic strength at the Purkinje cell to deep nuclear synapse. *Prog. Brain Res.* **124**, 257–273 (2000).
- Gaiarsa, J. L., Caillard, O. & Ben-Ari, Y. Long-term plasticity at GABAergic and glycinergic synapses: mechanisms and functional significance. *Trends Neurosci.* **25**, 564–570 (2002).
- Shapley, R. & Enroth-Cugell, C. in *Progress in Retinal Research* (eds Osborne, N. & Chader, G.) Vol. 3 263–346 (Pergamon, London, 1984).
- Baccus, S. A. & Meister, M. Retina versus cortex; contrast adaptation in parallel visual pathways. *Neuron* **42**, 5–7 (2004).
- Meister, M., Pine, J. & Baylor, D. A. Multi-neuronal signals from the retina: acquisition and analysis. *J. Neurosci. Methods* **51**, 95–106 (1994).
- Chichilnisky, E. J. A simple white noise analysis of neuronal light responses. *Network* **12**, 199–213 (2001).
- Cook, P. B., Lukasiewicz, P. D. & McReynolds, J. S. GABA_C receptors control adaptive changes in a glycinergic inhibitory pathway in salamander retina. *J. Neurosci.* **20**, 806–812 (2000).
- Schnitzer, M. J. & Meister, M. Multineuronal firing patterns in the signal from eye to brain. *Neuron* **37**, 499–511 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank members of the Meister laboratory, H. Sompolinsky and D. Fisher for advice. This work was supported by grants from the National Eye Institute (M.M. and S.A.B.) and the Human Frontier Science Program (T.H.).

Author Contributions T.H. and M.M. planned the study, T.H. and S.A.B. performed the experiments, and T.H. and M.M. completed the analysis and wrote the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.M. (meister@fas.harvard.edu).

ARTICLES

Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis

Lucas Pelkmans¹, Eugenio Fava², Hannes Grabner², Michael Hannus⁴, Bianca Habermann^{1,3}, Eberhard Krausz² & Marino Zerial¹

Endocytosis is a key cellular process, encompassing different entry routes and endocytic compartments. To what extent endocytosis is subjected to high-order regulation by the cellular signalling machinery remains unclear. Using high-throughput RNA interference and automated image analysis, we explored the function of human kinases in two principal types of endocytosis: clathrin- and caveolae/raft-mediated endocytosis. We monitored this through infection of vesicular stomatitis virus, simian virus 40 and transferrin trafficking, and also through cell proliferation and apoptosis assays. Here we show that a high number of kinases are involved in endocytosis, and that each endocytic route is regulated by a specific kinase subset. Notably, one group of kinases exerted opposite effects on the two endocytic routes, suggesting coordinate regulation. Our analysis demonstrates that signalling functions such as those controlling cell adhesion, growth and proliferation, are built into the machinery of endocytosis to a much higher degree than previously recognized.

Endocytosis supports a wide range of cellular functions, including nutrient uptake, downregulation of growth factors, cell-surface homeostasis, synaptic transmission and pathogen defence. Regulation of endocytic pathways is tightly coupled to the ability of cells to recognize, respond and adapt to extracellular stimuli. Different endocytic uptake mechanisms, including clathrin-mediated endocytosis (CME), caveolae/raft-mediated endocytosis, macropinocytosis and transactions between endosomes¹, are not only regulated by signalling molecules but also themselves contribute to signal transduction mechanisms^{2–4}. A detailed understanding of these multifaceted processes would benefit from a comprehensive analysis of the components necessary for endocytic transport and their responses to physiological stimuli, which is now possible owing to the availability of genomic sequences⁵. We reasoned that an important first step in understanding the supra-molecular regulation of endocytosis was to explore the role of the human kinome (the genomic collection of human protein, lipid and carbohydrate kinases) in endocytosis⁶. Only a small number of protein kinases have so far been implicated in endocytic transport^{1,7–9}. Here we undertook a genomic approach, systematically silencing predicted human kinases by RNA interference (RNAi), and determining the consequences of this silencing on two endocytic entry routes.

High-throughput viral entry screens

Endocytic transport is a multi-step process consisting of ligand binding and sequestration into internalization sites, the formation, transport, tethering and fusion of endocytic carriers with endosomal compartments, the sorting of cargo, and cargo distribution to its final destination^{1,10}. To identify genes regulating these events, we exploited two viruses that hijack the endocytic pathway to infect host cells^{11,12}, because virus infection is efficient, reproducible and easily measurable. Vesicular stomatitis virus (VSV) enters cells via

clathrin-mediated endocytosis into early and late endosomes¹³. Simian virus 40 (SV40) is transported to the endoplasmic reticulum (ER)¹⁴ upon caveolae-mediated endocytosis via caveosomes^{15–17}, as well as by non-caveolar, lipid raft-mediated endocytosis¹⁸. Notably, the same endocytic routes collectively regulate the signalling response of growth factor–receptor complexes^{4,19–21}.

The screening was performed using HeLa cells because their genes can be efficiently silenced with short interfering (si)RNA²². Three days after transfection (90% transfection efficiency and >70% reduction in messenger RNA), separate HeLa cell populations were infected with the two viruses using a low multiplicity of infection (MOI 0.1), to detect both reductions and increases in infection. As proof of principle (Supplementary Fig. S1a), we verified that VSV infection was specifically reduced by the silencing of established endocytic regulators like clathrin heavy chain and early endosome antigen 1 (EEA1; refs 23, 24), that SV40 infection was reduced by silencing of filamin and ceramide-glucosyl-transferase^{25,26}, and that both were reduced by ablation of *N*-ethylmaleimide sensitive factor (NSF)²⁷. The partial (40%) inhibition of SV40 infection observed upon silencing of caveolin-1 (Cav1) indeed reflects virus entry via a caveolin- and clathrin-independent, but lipid raft-dependent mechanism^{18,28}.

Genome-wide functional analysis of human kinases

The average relative infection index (RII, the ratio of infected siRNA-treated cells to infected control-treated cells) of VSV and SV40 was consistently reduced at least threefold by (1) silencing of the aforementioned genes, (2) using established inhibitors like GTPase-deficient dynamin2 (dynamin2-K44A), or (3) using methods such as endosomal pH neutralization (NH₄Cl), cytosol acidification (chlorpromazine) or cholesterol depletion (nystatin/progesterone) (RII ≤ 0.33, Fig. 1a and Supplementary Fig. S1a). We set this value

¹Max Planck Institute of Molecular Cell Biology and Genetics, ²MPI-CBG High-Throughput Technology Development Studio (HT-TDS), and ³Scionics Computer Innovation GmbH, Pfotenhauerstrasse 108, 01307 Dresden, Germany. ⁴Cenix Bioscience GmbH, Tatzberg 47, 01307 Dresden, Germany.

($RII \leq 0.33$) as the threshold for scoring inhibition, to identify necessary components or activators of the pathway. An RII greater than or equal to 3.00 was used as the threshold for scoring enhancement ($RII \geq 3.00$), indicating a repressor of the pathway. A kinase was assigned as being involved in both clathrin- and caveolae-mediated endocytosis pathways if both RIIs were <0.4 or >2.50 . Importantly, all RII values selected by these criteria fell outside the intervals defined by 3 s.d. of a large series of negative controls (Fig. 1a). See Supplementary Table S1 for a complete list of kinases and silencing phenotypes.

Out of the 590 human kinases screened (including lipid, protein, carbohydrate and hypothetical kinases), we identified a large number (208) as being involved in the two infectious entry routes. Whereas VSV infection was typically reduced, SV40 infection was enhanced for about half of the kinases. With the exception of CDK2

and GAK/auxillin, the kinases previously implicated in endocytosis (AAK1, Lck, c-Src, GSK3 β , protein kinase C (PKC) isoforms and CK2) were represented, showing that the screen identified the expected genes.

A number of control tests were used to verify that the high number of hits reflected a prominent regulatory role for the kinase in the infectious entry routes, rather than promiscuity of the assays. First, silencing of 50 randomly selected genes (see Supplementary Table S2) yielded only one hit (*OXCT1*, Fig. 1b), suggesting that the kinase is specifically enriched in regulators of the two infectious routes. Second, mRNA from 72% of the whole kinome and at least 90% of the hits could be detected in untreated HeLa cells, indicating that most of the kinase genes selected are expressed in HeLa cells. Third, we addressed kinome-wide the effects on cell proliferation and apoptosis (Fig. 1a). Although silencing of some genes reduced cell

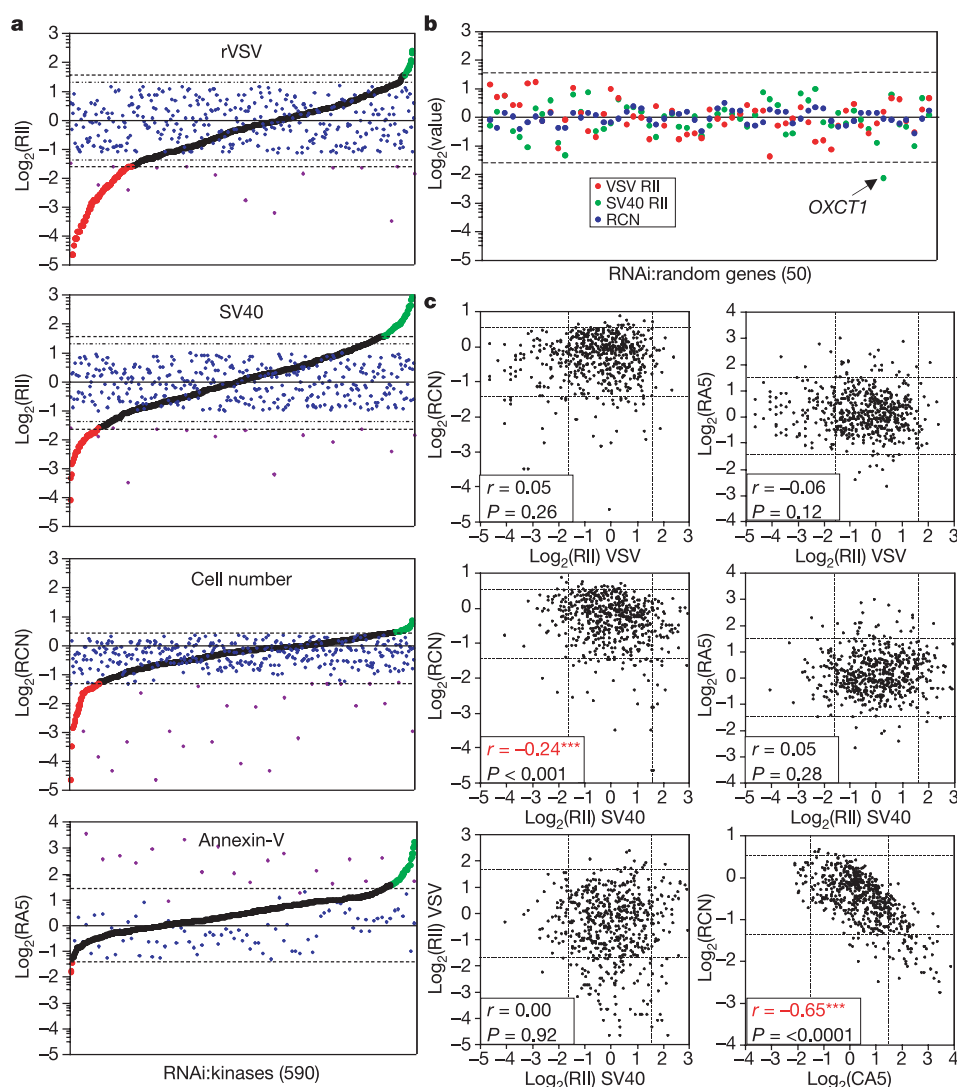


Figure 1 | High-throughput, genome-wide analysis of human kinases involved in infectious virus entry (VSV or SV40), cell proliferation and cell death. **a**, Average RII values, relative cell numbers (RCN) and relative Annexin-V staining (RA5), plotted on a log₂ scale, of the silencing of 590 kinases in triplicate. Relative Annexin-V staining calculated as the ratio of the total Annexin-V-positive fluorescence signal in the sample to signal in negative control wells. Kinases with values <0.33 (RIIs) or <0.4 (RCN and RA5) are shown in red. Kinases with values >3 (RIIs), >1.5 (RCN) or >2.5 (RA5) shown in green. Kinases with RII, RCN or RA5 values falling between the threshold limits are shown in black. Negative controls shown in blue and

positive controls in purple. Upper and lower dashed lines indicate scoring thresholds. **b**, Average RIIs and RCN (from triplicate samples) after silencing 50 randomly selected genes. Only one significant, SV40-specific hit (*OXCT1*; $RII = 0.23 \pm 0.11$) was observed. **c**, Correlation analysis between RII, RCN and RA5 values. Values are plotted as Log₂, dashed lines indicate scoring thresholds, and boxes show Spearman r and P values (two-tailed tests). A low but significant negative correlation (statistical significance indicated by asterisks) between SV40 and RCN was found, and an expected, strong negative correlation was observed between RCN and RA5.

proliferation and induced cell death and cytotoxicity (not shown), there was no positive correlation with virus infection. Quite the opposite was found for SV40, in that genes required for cell proliferation were suppressors of infection and vice versa, suggesting that the hits did not generally reflect toxicity or cell death. Fourth, more than 80% of the hits were pathway-specific. 92 kinases scored in VSV infection, 80 in SV40 infection, and of the 36 affecting both VSV

and SV40 infection, 23 gave inverse phenotypes, enhancing one pathway at the expense of the other.

Virus entry phenotypes include endocytosis phenotypes

Because our primary assays measure translation of an early-transcribed viral gene, some of the kinases identified could potentially be involved in virus-specific, post-endocytic events, including

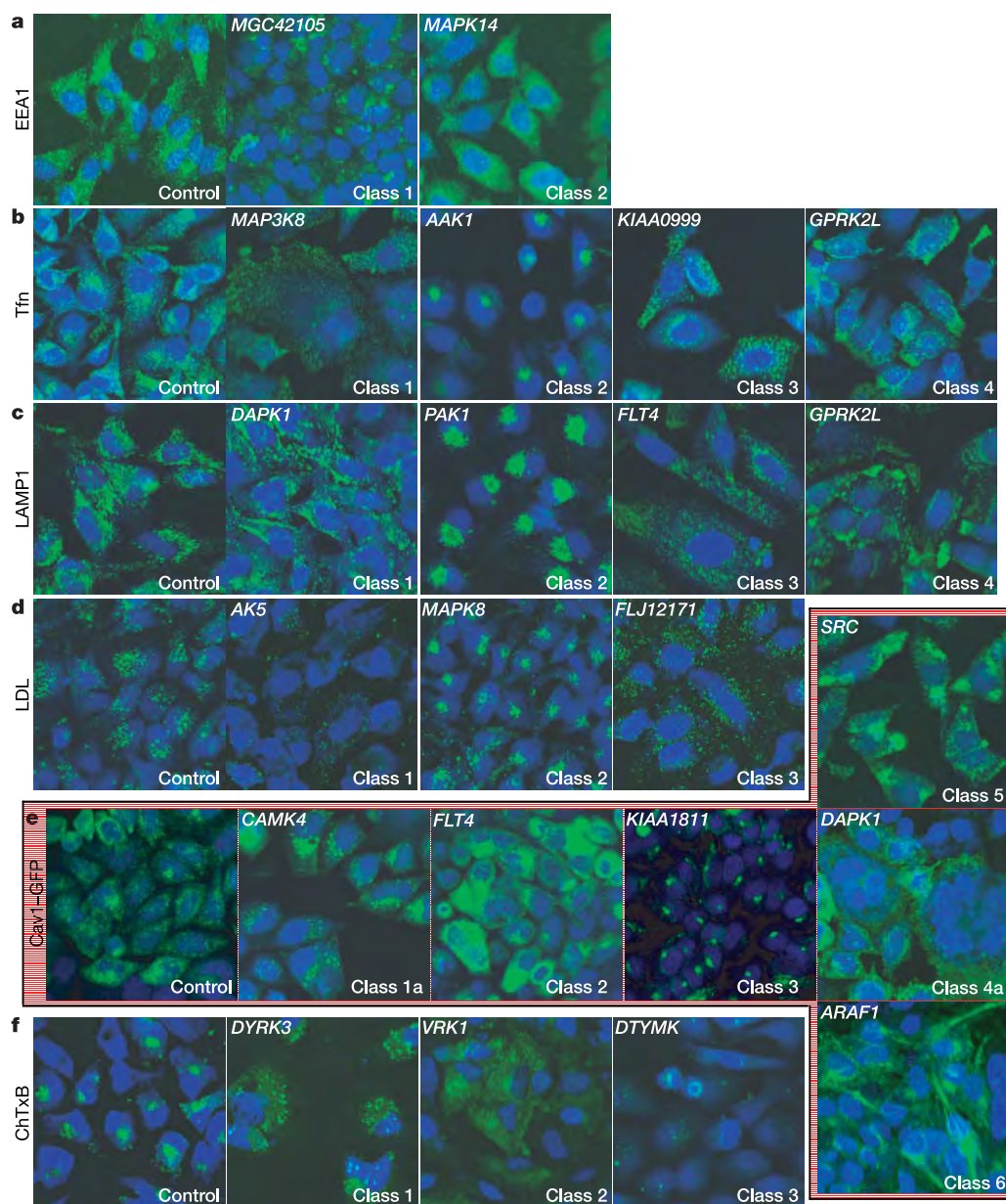


Figure 2 | Phenotypic profiling. **a–f**, Representative images of phenotype classes are shown. Phenotypes are a result of the silencing of the indicated kinase. Nuclei are pseudocoloured blue, endocytic staining is green. **a**, EEA1 immunostaining. Class 1, large vacuolar structures; Class 2, diffuse cytosolic staining. **b**, Fluorescent Tfn uptake (10 min). Class 1, accumulation in small spots in the cell periphery; Class 2, strong accumulation in the perinuclear area; Class 3, accumulation in enlarged cytoplasmic structures; Class 4, polarized accumulation in plasma membrane protrusions. **c**, LAMP1 immunostaining. Class 1, structures aligning at the plasma membrane; Class 2, strong and condensed perinuclear accumulation; Class 3, distribution of structures throughout the cytoplasm; Class 4, polarized accumulation in plasma membrane protrusions. **d**, Fluorescent LDL uptake (30 min). Class 1, strongly reduced signal, accumulated in structures distributed throughout

the cytoplasm; Class 2, strong and condensed perinuclear accumulation; Class 3, accumulation in vesicular structures throughout the cell.

e, Cav1–GFP staining (images highlighted in pink). Class 1, signal in vesicular structures distributed throughout the cell (Class 1a) or more perinuclear (Class 1b, not shown); Class 2, diffuse staining distributed over the whole cell; Class 3, accumulation in a condensed perinuclear spot; Class 4, small punctuate structures and diffuse staining distributed on the cell surface (Class 4a) or in straight lines at the cell surface (Class 4b, not shown); Class 5, accumulation in peripheral foci; Class 6, diffuse surface staining and intracellular tubules. **f**, Fluorescent ChTxB uptake (60 min). Class 1, accumulation in vesicular structures distributed throughout the cytosol; Class 2, peripheral and diffuse staining; Class 3, strongly reduced signal.

synthesis and delivery of virus receptors, and transcription/translation of viral DNA/RNA. To validate the role of kinases in endocytosis, we applied secondary, high-content assays using an automated, high-throughput spinning disk confocal microscope (see Methods). First, we re-screened the complete kinome for fluorescent transferrin (Tfn) uptake and trafficking, and monitored four different phenotypic classes (Tfn Classes 1–4; see Fig. 2 and Supplementary Table S1). For a subset of 50 kinases (selected for their inverse phenotypes or unknown function), we monitored an additional 18 phenotypic classes for the two endocytic routes (Fig. 2). These included the morphology and distribution of cargo in the early endosome (stained using antibodies against EEA1) and late endosome (using antibodies against lysosomal antigen marker protein 1 or LAMP1), the distribution and morphology of caveolin-1-GFP (Cav1-GFP)-positive structures, trafficking of cargo to late endosomes (using fluorescent low density lipoprotein, LDL), and trafficking of cargo by caveolae/raft-mediated endocytosis (using fluorescent cholera toxin B, ChTxB) (Fig. 2 and Supplementary Table S3).

This phenotypic profiling verified that at least 72% of kinases

affecting VSV infection (92/128) also played a role in CME, and that 87% of the phenotypically profiled kinases involved in SV40 infection (34/39) yielded perturbations in caveolae/raft trafficking. By extrapolation, the majority of kinases identified in the primary virus infection screens are thus predicted to be involved in endocytic trafficking. The secondary assays also confirmed that the two endocytic routes are differentially regulated: 21% (27/129) of the kinases acting in CME affected SV40 infection, and none of the kinases involved in caveolae/raft-mediated endocytosis affected VSV infection.

To order the kinases into a functional pattern, we performed a two-step cluster analysis²⁹ (Supplementary Table S3). First, hierarchical clustering was performed on the VSV/SV40 RIIs, differentiating ten groups of kinases with highly correlating phenotypes (Fig. 3, Groups 1–10). Second, within each group, kinases were clustered according to the RIIs and all phenotypic classes of the endocytic routes shown in Fig. 2 (Fig. 3). In addition, we manually annotated structural properties and previously ascribed functions (Supplementary Table S3 and references therein), and hierarchically clustered those kinases with well-established functions within

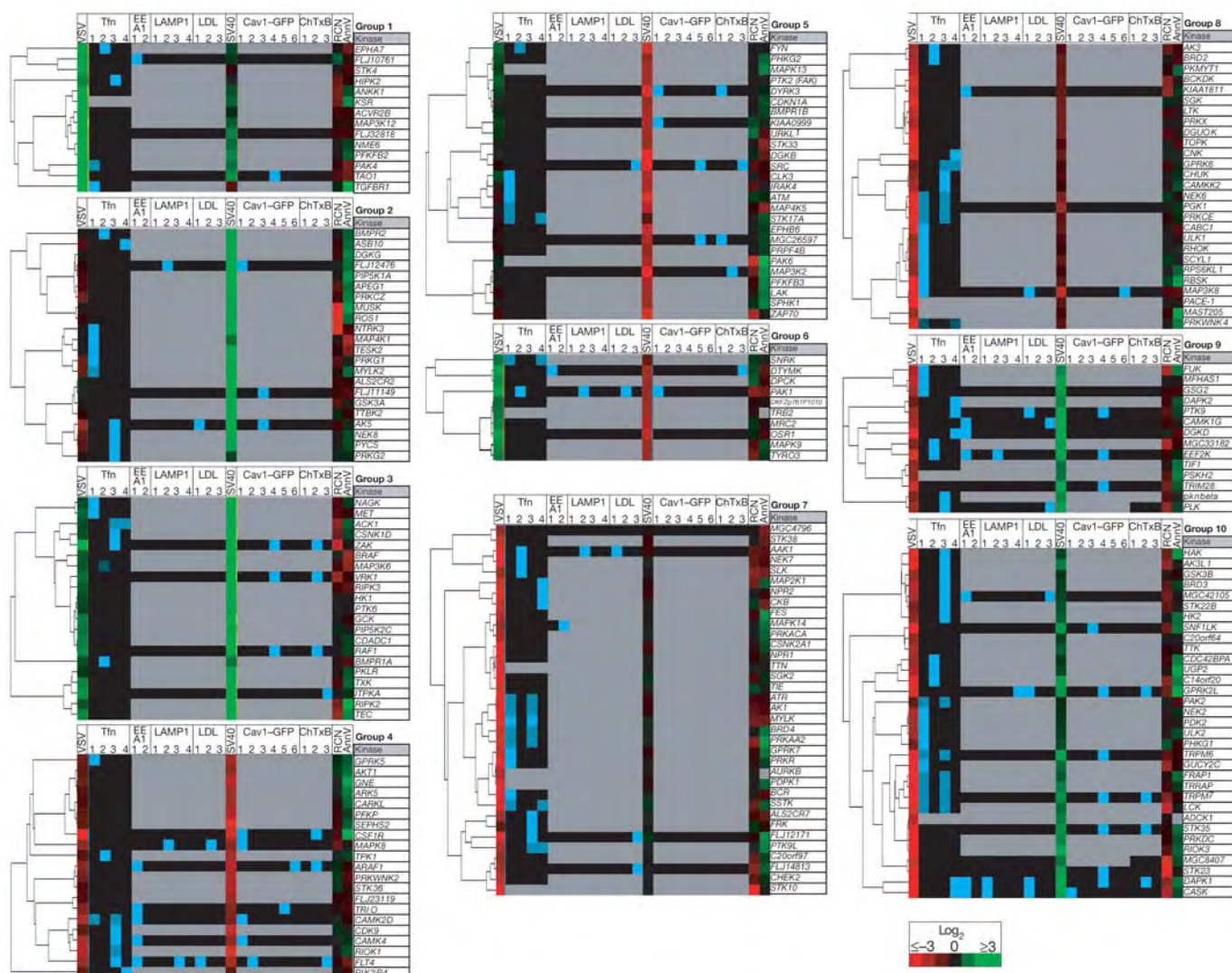


Figure 3 | Hierarchical two-step clustering of RIIs and phenotypic profiles. For a full explanation, see Supplementary Table S3. First, the SV40 and VSV RIIs were hierarchically clustered²⁹. Ten groups that clustered strongly (>0.9) were identified. Next, the RIIs, RCN, RA5 and phenotypic profiles (according to the Classes outlined in Fig. 2) of the kinases were hierarchically

clustered within each group, resulting in the depicted list of kinases. For the manual annotation of previously ascribed functions for each kinase, see Supplementary Table S3. NCBI URLs for each kinase are listed in Supplementary Table S1.

these different functional groups (Fig. 4). Cumulatively, the results revealed the existence of interconnected functional groups, as illustrated in the following examples.

Regulation by metabolic kinases

Silencing of kinases regulating metabolic pathways might be expected to yield unspecific effects. Notably, 17 metabolic kinases scored specifically in VSV and 14 in SV40 infection. This specificity might in part reflect the need for specific lipids or glycans in both

endocytic pathways. For example, *GNE* (Fig. 3, Group 4) regulates the production of sialic acid, which is required for biosynthesis of the SV40 receptor GM1 (ref. 26), and silencing of *GNE* specifically reduced SV40 infection. Knock-down of *FLJ12171* (also known as *FN3KRP*; Fig. 3, Group 7), a kinase that removes fructosamines from proteins, specifically blocked VSV infection and resulted in Tfn accumulation in enlarged endosomes (Tfn Class 3). This suggests unexpected regulatory functions of fructosamine modifications in the early endocytic pathway.

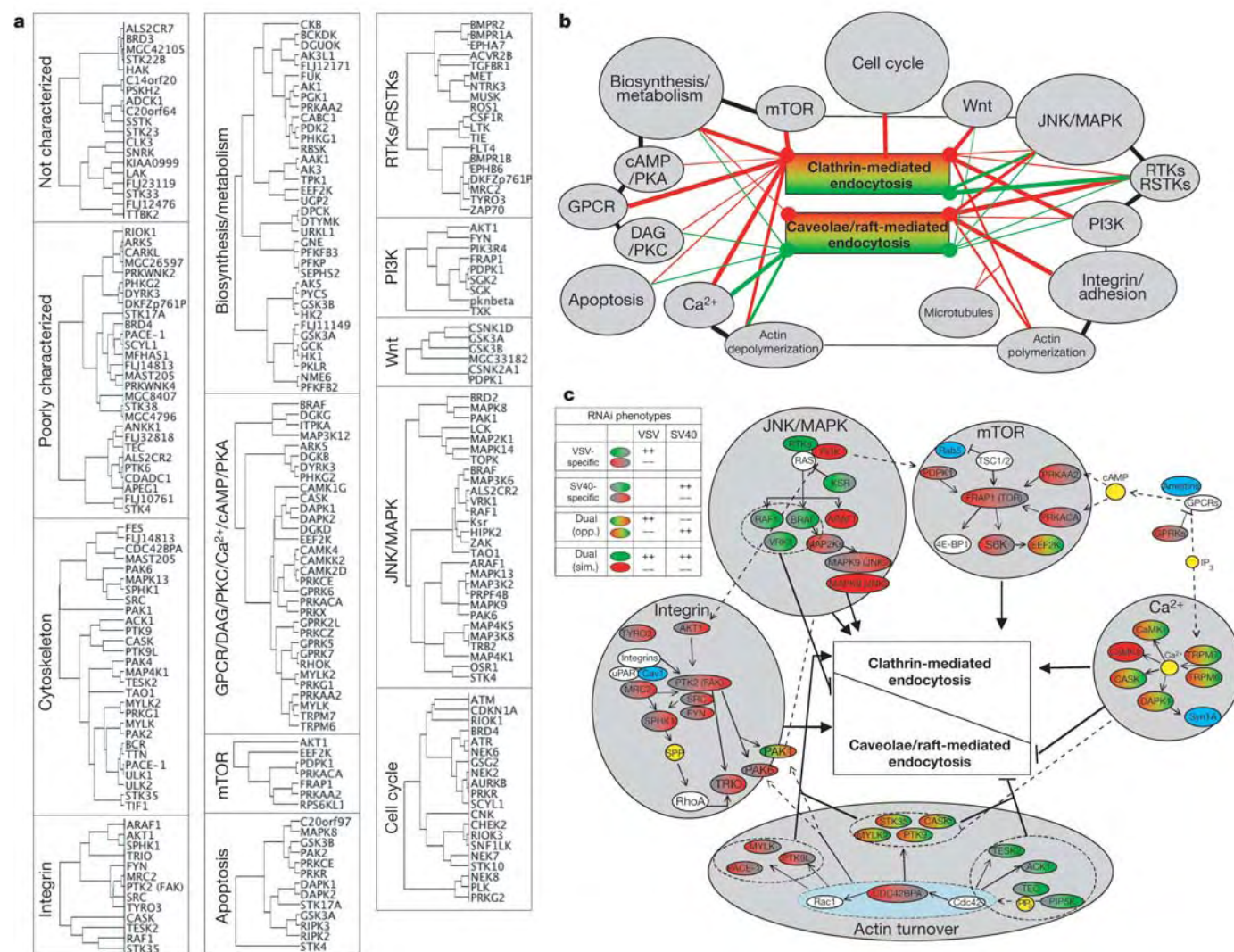


Figure 4 | Effects of signalling pathways on endocytosis. **a**, Using the previously ascribed functions (see Supplementary Table S2), kinases were assigned to 11 groups, in a non-exclusive manner: (1) not characterized at all, (2) poorly characterized, or involved in (3) regulation of the cytoskeleton, (4) integrin signalling and adhesion, (5) biosynthesis and metabolism, (6) GPCR signalling cascades involving DAG, Ca²⁺ and cAMP, and the kinases PKC and PKA, (7) mTOR signalling, (8) apoptosis, (9) signalling from RTKs and RSTKs, (10) kinases functionally linked to PI(3)K, (11) the Wnt signalling pathway, (12) the JNK/MAPK cascade, and (13) the cell cycle. Kinases were then phenotypically clustered within each group (see Supplementary Fig. S2). **b**, The general effects on each endocytic route of the functional groups (see Supplementary Fig. S2) were deduced and indicated by coloured lines. Note that the colour code corresponds to the RNAi phenotype of a particular kinase. Red indicates that a kinase is required or is an activator of the pathway, and green indicates a kinase that is a suppressor of the endocytic pathway. **c**, Four representative examples of highly interconnected signalling networks. The TOR pathway regulating cell

growth in response to nutrients specifically maintains or even stimulates CME. The integrin-signalling cascade, which controls focal adhesion assembly and turnover, is specifically required to activate or maintain caveolae/raft-mediated endocytosis. Ca²⁺/calmodulin-dependent signalling and actin turnover orchestrated by Cdc42/Rac1 are two interconnected systems that exert opposite effects on the endocytic routes, suppressing caveolae/raft-mediated endocytosis while sustaining CME. A more complex regulatory activity is observed for the stress-induced and mitogenic signalling (JNK/MAPK) cascades, which can exert positive as well as negative effects on both endocytic routes. There are also connections between endocytosis and the cell cycle as well as programmed cell death. The phenotypes of the individual kinases are indicated by colours and by the position of that colour within the kinase symbol (colour on the left, VSV phenotype; colour on the right, SV40 phenotype). Generalized effects of networks on the endocytic routes are indicated with black arrows (required or activating) or black inhibitory symbols (suppressing).

VSV infection and CME

Of the 80 kinases for which silencing strongly blocked VSV infection with little effect on SV40 infection (Fig. 3, Groups 7 and 8), 48 acted at different stages of the endocytic pathway. Phenotypes were observed in uptake, recycling and degradative transport. For example, silencing of *SSTK* (also called *TSSK6*) a Ser/Thr kinase with unknown function, inhibited internalization of Tfn (Tfn Class 1). Silencing of *AAK1*, which negatively regulates transferrin endocytosis, increased transport and accumulation of Tfn in perinuclear endosomes (Tfn Class 2), but also reduced VSV infection and uptake of low-density lipoprotein (LDL) (LDL Class 1).

Regulation by TOR-dependent signalling. Tor2p has been shown to regulate α -mating-factor receptor endocytosis in *S. cerevisiae*. Silencing of *FRAP1* (mammalian TOR) and five members of the TOR signalling pathway (*PDPK1*, *PRKAA2*, *PRKACA*, *RPS6KL1*, *EEF2K*) (Fig. 3, Groups 7–10 and Fig. 4) specifically blocked VSV infection, Tfn internalization (*RPS6KL1*, Tfn Class 1), intracellular trafficking (*PRKAA2*, *FRAP1*, Tfn Class 1 and 3), or recycling (*EEF2K*, Tfn Class 2). This shows that the nutrient-dependent TOR signalling pathway not only regulates ligand uptake but also imposes control at multiple steps of the CME pathway in human cells.

Regulation by G-protein receptor-linked kinases. Phosphorylation of G-protein-coupled receptors (GPCRs) by G-protein receptor-linked kinases (GPRKs) leads to their arrestin-dependent down-regulation by CME. We found that silencing of four different GPRKs expressed in HeLa cells (*GPRK5*, *GPRK6*, *GPRK7* and *RHOK/GRK1*) specifically reduced VSV infection and Tfn internalization (Tfn Class 1 and 4). GPRKs can therefore exert a more general form of control over CME, beyond the internalization of GPCRs.

Cytoskeleton-dependent transport and cell polarity. Our screen assigned a prominent role in endocytosis to a number of regulators of both the actin and tubulin cytoskeleton. Silencing of *MYLK* (Fig. 3, Group 7), which phosphorylates myosin light chain, caused an early block in Tfn uptake (Tfn Class 1 and 3). Silencing of *FLJ14813* (also called *MASTL*), a microtubule-associated Ser/Thr (MAST)-like kinase, reduced the accumulation of LDL in perinuclear structures (LDL Class 3), suggesting that it is required for late endocytic transport.

Ablation of *PTK9L* (Fig. 3, Group 7), which binds monomeric actin and can be phosphorylated by c-Src, CK2 (*CSNK2A1*, see below) and PKC ζ (*PRKCZ*), resulted in redistribution of Tfn-positive vesicles to membrane protrusions and lamellae (Tfn Class 3 and 4). PKC ζ , which regulates cell polarization, might therefore re-programme the activity and distribution of the recycling system via *PTK9L*. Silencing of *ULK1* and *ULK2* blocked Tfn internalization (Tfn Class 1) (Fig. 3, Groups 7, 8 and 10). Both of these kinases are homologous to *C. elegans* Unc51.1, which binds to microtubules, regulates Rab5-dependent endocytic transport³⁰ (as part of a complex with SynGAP and syntenin), and is involved in axonal elongation.

SV40 infection and caveolae/raft-mediated endocytosis

Because SV40 can use caveolae-mediated endocytosis, as well as a very related (if not identical) uptake route internalizing lipid rafts in the absence of Cav1 (ref. 18), the kinases we identified as affecting SV40 infection might be involved in both endocytic routes. Indeed, silencing of ten kinases in HeLa cells conditionally depleted for Cav1 by RNAi affected SV40 infection as in control cells (Supplementary Fig. S1b). These data suggest that this group of kinases also function in caveolae-independent, raft-dependent endocytosis, acting as regulators of this enigmatic transport route.

Among the 43 kinases specifically required for SV40 infection, we detected a series of interesting phenotypes to do with caveolae trafficking. For example, silencing of *KIAA0999* (*FLJ12240*), an unknown Ser/Thr kinase, and *DYRK3*, a MAPK-related kinase involved in cell growth and survival (Fig. 3, Group 5), caused a redistribution of green fluorescent protein (GFP)-labelled Cav1 and

internalized ChTxB to enlarged vesicular structures (ChTxB Class 1). We also observed clustering of caveolae at peripheral foci (Cav1–GFP Class 5), tubulation of intracellular Cav1–GFP-positive structures (Cav1–GFP Class 6), and diffuse cell-surface staining (Cav1–GFP Class 4) (Fig. 2). Using quantitative total internal reflection fluorescence microscopy, we could show a specific role for kinases at different stages of the caveolar assembly and transport cycle³¹.

Regulation by focal adhesion and integrin signalling. SV40 infection was specifically reduced by silencing of (1) *PTK2/FAK* and *SRC* (Fig. 3, Group 5), which regulate integrin-dependent focal adhesion assembly and turnover, (2) *MRC2*, a collagen receptor tyrosine kinase, and (3) several other members of cell adhesion-dependent signalling (*FYN*, *AKT1*, *TYRO3*, *PAK6* and *SPHK1*, Fig. 3). The clustering of Cav1–GFP-positive and ChTxB-positive structures at peripheral foci (Cav1–GFP Class 5) seen after silencing of *SRC* and *MGC26597* (a predicted protein with potential phosphatidylinositol-4-phosphate-5-kinase, PI(4,5)K activity), suggests that these kinases regulate caveolae/raft-mediated traffic to, and from, focal adhesions, supporting the idea that the extracellular matrix can modulate this trafficking route^{32,33}.

Cdc42-mediated suppression of endocytosis. As a prerequisite to internalization by caveolae, SV40 is thought to stimulate a cycle of cortical actin depolymerization and polymerization³⁴. Interestingly, among the hits that specifically enhanced SV40 infection were two PI(4,5)Ks (*PIP5K1A*, *PIP5K2C*), a PH-domain-containing Src-family kinase (*TEC*) that activates Cdc42 and actin polymerization, and two Cdc42 effectors, *ACK1/TNK2* and *TESK2*, which enhance actin polymerization. These data suggest that caveolae/raft-mediated endocytosis is suppressed by the stabilizing effect that Cdc42- and PI(4,5)P₂-dependent signalling have on cortical actin. However, as *MGC26597* is required for SV40 infection, different PI(4,5)Ks might regulate the actin cycle and exert opposite effects on this endocytic route.

Coordination between endocytic pathways

Among the 36 kinases found to have a role in both VSV and SV40 infection, a subset (23 kinases) exerted opposite effects on the two virus entry mechanisms. For those kinases for which silencing inhibited VSV infection and enhanced SV40 infection (Fig. 3, Groups 9 and 10), phenotypic profiling revealed a functional association with transport to late endosomes. For example, knockdown of two Ca²⁺ channels with kinase domains (*TRPM6*, *TRPM7*) inhibited VSV infection and stimulated SV40 infection. Similar effects were observed for several Ca²⁺/calmodulin-activated kinases (for example, *CASK*, *DAPK1*, *DAPK2*, *DGKD*), silencing of which caused early or late endosomes to accumulate underneath the cell surface. The concomitant enhancement of SV40 infection suggests either that caveolae/raft-mediated endocytosis compensates for reduced CME, or that cargo internalized by caveolae bypasses the endosomes, leading to increased transport of SV40 to the ER and enhancing infection²⁸.

Actin cytoskeleton. Among the kinases coordinating the two endocytic routes were effectors of Cdc42 and Rac1, implicated in the regulation of the actin cytoskeleton (see above). Their activity in endocytosis correlated with the intracellular distribution of endosomes, as shown by the opposite effects of *PTK9* and *PAK1*. Silencing of *PTK9* (twinfilin), which binds monomeric actin and is targeted by Cdc42 and Rac1 to regions of high actin turnover, inhibited VSV and enhanced SV40 infection, and redistributed endocytic organelles to the cell periphery. Conversely, depletion of the p21-activated kinase *PAK1*, an effector of Cdc42 and Rac1 that depolymerizes actin, increased VSV infection, reduced SV40 infection, and induced the clustering of endocytic organelles in the perinuclear region.

Wnt/ β -catenin signalling. Silencing of two glycogen synthase kinases (*GSK3A* and *GSK3B*) and three casein kinases (*CSNK1D*, *MGC33182/CSNK1A1L* and *CSNK2A1*), all known to signal for proteasomal degradation of β -catenin in the absence of Wnt ligand,

reduced VSV and enhanced SV40 infection. Notably, ablation of *GSK3B* and *CSNK1D* resulted in accumulation of Tfn in early endosomes (Class 3), suggesting that β -catenin degradation is regulated by the same kinases that control endocytic downregulation of cell-surface components (for example, cadherins).

Mitogenic and other signalling pathways. Our comparative analysis

of phenotypes in clathrin- versus caveolar/raft-dependent pathways, cell growth, proliferation and apoptosis has provided an integrated view of signalling and endocytosis. The screen identified pro-apoptotic kinases (for example, Death-associated protein kinase 1 DAPK1), the ablation of which yielded an endocytic phenotype (Fig. 4). One surprise was the extent to which the mitogenic signalling cascade participates in endocytosis (Fig. 4): nine receptor tyrosine kinases (RTKs) and four receptor Ser/Thr kinases (RSTKs) expressed in HeLa cells were either required for (four for SV40, none for VSV infection alone, two for both SV40 and VSV infection) or suppressed (five for SV40 and three for VSV infection) the virus infectious entry routes. In addition, we identified at least 28 kinases of the mitogenic signalling/JNK cascade^{35–37} with a variety of different phenotypic classes. Among the silencing phenotypes of the 37 suppressors of SV40 infection (Fig. 3, Group 3, for example, *ZAK*, *VRK1* and *RAF1*), we frequently observed diffuse or fine, punctuate staining of Cav1–GFP in the periphery of the cell (Cav1–GFP Class 4), with ChTxB accumulating in enlarged intracellular vesicles (ChTxB Class 2).

As CME and endosome fusion are arrested during mitosis^{38,39}, we were not surprised to see the screen identify kinases regulating cell-cycle progression (Fig. 4). Silencing of *MAP2K1* (*MEK1*) and *CNK* (also *PLK3* or Polo-like kinase-3), known to fragment the Golgi during the cell cycle, re-targeted Tfn-containing endosomes to plasma membrane protrusions (Tfn Class 4), suggesting that this signalling cascade also regulates the distribution of recycling endosomes. This mechanism might serve to partition endosomes between daughter cells or supply membranes to the cleavage furrow during cytokinesis.

Changes in Cav1 and MAPK p38 phosphorylation

To explore the function of kinases on the target endocytic machinery, we inspected the phosphorylation of Cav1 because it has been suggested to regulate endocytosis of caveolae/lipid rafts^{40,41}. Unexpectedly, we detected increased levels of phosphorylated Cav1 and changes in localization upon silencing of several kinases required for SV40 infection (Fig. 5a). Because silencing of *ARAF/ARAF1* and *DYRK3* also modified the cellular distribution of Cav1–GFP (Cav1–GFP Class 1 and 6), coat stability and dynamics of caveolae on the cell surface³¹, our data suggest that, in addition to Cav1 phosphorylation, a network of kinases positively and negatively regulates the dynamics and transport activity of caveolae.

We observed cytosolic staining of EEA1 (EEA1 Class 2) upon silencing of the stress-activated kinase *MAPK14* (p38 β ; Fig. 2). Like JNK, p38 MAPK lies at the centre of stress-induced signalling cascades³⁷ and regulates endocytosis by phosphorylating components of the endocytic machinery^{42,51}. We tested whether kinases regulating CME could act through phosphorylation of p38. Phosphorylated p38 was low in non-stimulated cells, but increased upon silencing of several kinases required for VSV infection (Fig. 5b). Concomitantly, and although the non-phosphorylated kinase is mainly cytosolic⁴², phosphorylated p38 associated with enlarged endosomes (Fig. 5b). Thus, ablation of various kinases mimics the stress-induced activation of p38 MAPK³⁷ resulting in recruitment of phosphorylated p38 to endosomes and modulation of endocytosis.

Discussion

Since the initial observations that kinases regulate endocytosis⁴³, a more general appreciation of their role at distinct steps of the endocytic pathway has been lacking¹. Our genome-wide screen fills this gap by contributing a comprehensive analysis of the involvement of human kinases in two endocytic routes. For example, the identification of new kinases regulating the stability of the caveolar coat and caveolae dynamics provides long-sought-after regulators of this transport route³¹. Furthermore, our screen reveals some new design principles governing endocytosis. First, we identified an unexpected number of kinases (210) involved in endocytosis, several of which

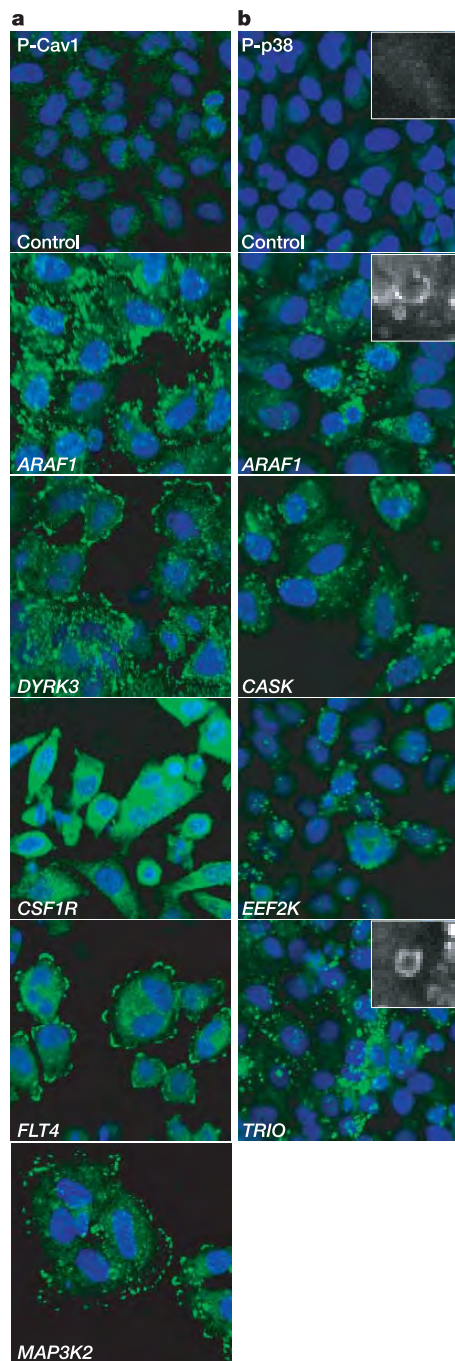


Figure 5 | Phosphorylation of Cav1 and p38 MAPK. **a**, HeLa cells silenced for the indicated kinases were fixed and stained with a monoclonal antibody against caveolin-1 phosphorylated on Tyr 14 (P-Cav1). In control cells, some P-Cav1 is detected but the level is low. Silencing of the indicated kinases induces hyperphosphorylation and location-specific phosphorylation of Cav1. **b**, HeLa cells silenced for the indicated kinases were fixed and stained with a monoclonal antibody against phosphorylated p38 (P-p38). In control cells, low levels of phosphorylated p38 are detected. Silencing of the indicated kinases induces hyperphosphorylation of P-p38, which is then detected on intracellular, enlarged endosome-like structures (see insets).

(47) were poorly characterized or uncharacterized. Although not all kinases would be expected to play a direct mechanistic role in endocytosis (see below), we observed high selectivity for one of the two endocytic routes, even for regulators of metabolic functions. Second, the finding that some kinases can positively affect one endocytic route at the expense of the other suggests that clathrin- and caveolae-mediated endocytosis are coordinated to ensure a proper balance in membrane homeostasis and signalling activities. Third, a large group of the kinases identified in our endocytosis screen also function in various signalling pathways, strengthening the emerging paradigm that endocytic transport and signal transduction are tightly coupled. This might in part explain the complexity observed, in that only some kinases within a given signalling network act directly upon the endocytic machinery, with the majority having only a modulatory role. A generalization of the higher-order control of signalling cascades on the endocytic pathway is depicted in Fig. 4a, b. Depending on the signalling system, kinases entrain the endocytic machinery to respond with either clathrin- or caveolae/raft-dependent endocytosis, or both. It appears that endocytosis is subjected to regulatory control by the signalling machinery even under normal growth conditions. For example, our data and others⁵¹ argue for a role of the stress-response p38MAPK⁴² in endocytosis also under physiological conditions.

These observations form a framework for rethinking how cells respond and adapt to different stimuli. Endocytosis has a much wider role than downregulating activated growth factor–receptor complexes^{44–47}, by providing spatio-temporal regulation to the signalling response. First, growth factor–receptor complexes can modify their own intracellular itinerary, by activating downstream kinase cascades. For example, four mitogenic kinases suppress caveolae/raft-mediated endocytosis by regulating Cav1 phosphorylation. As caveolae keep signalling molecules in an inhibited state⁴⁸, this feedback mechanism might be universally applied in controlling cell proliferation: A significant correlation across the kinome (Fig. 2c) indicates that kinases required for cell proliferation tend to suppress caveolae/raft-mediated endocytosis, whereas inhibitory kinases have the opposite effect. In addition to these general considerations, the endocytic response to kinases clearly also depends on the cell type. Second, by regulating the motility of endosomes along the cytoskeleton, kinases might enable signalling molecules on endosomes to propagate signals from the plasma membrane to the nucleus by direct transport rather than simple diffusion⁴⁹. Our findings have also implications for biomedical research and drug development, through the identification of new mechanisms of action and potential therapeutic targets. Finally, by extending our high-throughput analysis to the entire human genome, we will eventually be able to obtain a functional genetic map of endocytic pathways. The challenge then will be to translate this framework into a quantitative model that incorporates parameters including transport kinetics, changes in morphology and localization, and response magnitude, and can predict responses under different physiological conditions.

METHODS

siRNA design and silencing efficiency. Three unique sites within the coding sequence of each human kinase were selected (Cenix Bioscience GmbH), 21-mer oligonucleotides synthesized (Ambion Inc.) and gene silencing potential verified by real-time polymerase chain reaction (PCR). The most potent 21-mers (resulting in >70% reduction of mRNA levels after 3 days) were included in the kinase siRNA library. Re-screening of 50 kinases using the VSV and SV40 infectious entry assays with independent siRNA sequences resulted in the same phenotype in 48/50 (SV40) and 50/50 cases (VSV).

High-throughput siRNA screening of infectious virus entry. rVSV–GFP and SV40 (strain 776) were prepared according to published protocols^{17,50}. Between cell plating and fixation, no washing steps were required. HeLa cells (4,000 per well) were plated in 80 µl of complete medium in 96-well plates and incubated at 37 °C and 5% CO₂. The next day, 1 µl of 10 µM siRNA, diluted in 16 µl of OptiMax with 0.4 µl oligofectamine in 2.6 µl OptiMax, was added directly to the cells and incubated for 3 days. After 3 days, cells were incubated for 3 h with

rVSV–GFP or for 36 h with SV40 at 37 °C and 5% CO₂. After infection, cells were fixed with paraformaldehyde (PFA), permeabilized with Triton X-100 (TX100) and stained with 4,6-diamidino-2-phenylindole (DAPI). A monoclonal antibody directly conjugated to Alexa Fluor 488 (AF488) was used to detect SV40 large T-antigen expression. In separate plates, gene-silenced, PFA-fixed and BSA-blocked cells were stained with 1 µg ml^{−1} AF488–Annexin-V for 1 h in PBS.

Automated imaging and quantification. In a fully automated and double-blind manner, five images were taken per well, the total number of cells (DAPI) and infected cells (GFP/AF488) per well calculated, and average and standard deviations for triplicate experiments determined. These were used to calculate RII, relative cell number (RCN) and relative Annexin-V staining (RA5) (Fig. 1). For determination of the signal noise, numbers obtained from each field of control-treated cells were not averaged, but regarded as individual measurements of the noise spectrum.

High-content endocytosis assays. Fixed and permeabilized cells were stained with anti-EEA1 or -LAMP1 antibodies (BD Biosciences Pharmingen) and appropriate secondary antibodies. Cav1 distribution was studied in siRNA-treated HeLa cells stably expressing Cav1–GFP²⁸. For ligand trafficking, siRNA-treated HeLa cells were incubated in 1 µg ml^{−1} AF488–Tfn for 10 min, 1 µg ml^{−1} Dil–LDL for 30 min, and 1 ng ml^{−1} AF488–ChTx for 60 min. Cells were washed, fixed and stained with DRAQ5 (Alexis Biochemicals) to visualize nuclei. The Opera automated spinning-disk confocal microscope (Evotec Technologies GmbH) was used to image 96-well plates, using 488-nm, 532-nm and 633-nm laser lines and a water-immersion 20 × objective with a numerical aperture of 1.2. Twenty fields per well (each containing ~100 cells under control conditions) were imaged and experiments were performed in triplicate. Images were analysed manually and phenotypes scored when appearing in at least five fields in each triplicate.

Hierarchical clustering. The phenotypic classes were first converted into a numerical representation: The value 1.000 (2⁰) was assigned to all samples with no phenotypes, and the value 0.125 (2^{−3}) to the phenotype observed. If two phenotypes were observed within one assay (for example, Tfn Class 1 and 3), each phenotype was given the value 0.250 (2^{−2}). In a few cases where the phenotype was less penetrant, it was given the value 0.500 (2^{−1}). For Fig. 3, the RII values were loaded into Cluster 3.0, log₂-transformed and hierarchically clustered using uncentred correlation and centroid linkage²⁹, yielding ten infection-phenotype groups of kinases clustering >0.9. Next, the RII, RCN, RA5 and phenotypic profiling values were loaded into Cluster 3.0 by infection-phenotype group, and processed as above. In Fig. 3, 11 functional groups were assembled on the basis of functions previously ascribed to the kinases. Next, the RII, RCN, RA5 and phenotypic profiling values were loaded into Cluster 3.0 by functional group, and processed as above. Java TreeView was used to make cluster trees and graphic displays of the phenotypes.

Received 28 January; accepted 17 March 2005

Published online 11 May 2005.

- Conner, S. D. & Schmid, S. L. Regulated portals of entry into the cell. *Nature* **422**, 37–44 (2003).
- Di Fiore, P. P. & De Camilli, P. Endocytosis and signaling: an inseparable partnership. *Cell* **106**, 1–4 (2001).
- Gonzalez-Gaitan, M. Signal dispersal and transduction through the endocytic pathway. *Nature Rev. Mol. Cell Biol.* **4**, 213–224 (2003).
- Miaczynska, M., Pelkmans, L. & Zerial, M. Not just a sink: endosomes in control of signal transduction. *Curr. Opin. Cell Biol.* **16**, 400–406 (2004).
- Carpenter, A. E. & Sabatini, D. M. Systematic genome-wide screens of gene function. *Nature Rev. Genet.* **5**, 11–22 (2004).
- Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912–1934 (2002).
- Lucocq, J., Warren, G. & Pryde, J. Okadaic acid induces Golgi apparatus fragmentation and arrest of intracellular transport. *J. Cell Sci.* **100**, 753–759 (1991).
- Simonsen, A., Wurmser, A. E., Emr, S. D. & Stenmark, H. The role of phosphoinositides in membrane transport. *Curr. Opin. Cell Biol.* **13**, 485–492 (2001).
- Nguyen, C. & Bibb, J. A. Cdk5 and the mystery of synaptic vesicle endocytosis. *J. Cell Biol.* **163**, 697–699 (2003).
- Mellman, I. Endocytosis and molecular sorting. *Annu. Rev. Cell Dev. Biol.* **12**, 575–625 (1996).
- Helenius, A., Kartenbeck, J., Simons, K. & Fries, E. On the entry of Semliki forest virus into BHK-21 cells. *J. Cell Biol.* **84**, 404–420 (1980).
- Smith, A. E. & Helenius, A. How viruses enter animal cells. *Science* **304**, 237–242 (2004).
- Sieczkarski, S. B. & Whittaker, G. R. Differential requirements of rab5 and rab7 for endocytosis of influenza and other enveloped viruses. *Traffic* **4**, 333–343 (2003).

14. Kartenbeck, J., Stukenbrok, H. & Helenius, A. Endocytosis of simian virus 40 into the endoplasmic reticulum. *J. Cell Biol.* **109**, 2721–2729 (1989).
15. Anderson, H. A., Chen, Y. & Norkin, L. C. Bound simian virus 40 translocates to caveolin-enriched membrane domains, and its entry is inhibited by drugs that selectively disrupt caveolae. *Mol. Biol. Cell* **7**, 1825–1834 (1996).
16. Stang, E., Kartenbeck, J. & Parton, R. G. Major histocompatibility complex class I molecules mediate association of SV40 with caveolae. *Mol. Biol. Cell* **8**, 47–57 (1997).
17. Pelkmans, L., Kartenbeck, J. & Helenius, A. Caveolar endocytosis of simian virus 40 reveals a new two-step vesicular-transport pathway to the ER. *Nature Cell Biol.* **3**, 473–483 (2001).
18. Damm, E.-M. *et al.* Clathrin and caveolin-1 independent endocytosis: entry of simian virus 40 into cells devoid of caveolae. *J. Cell Biol.* **168**, 477–488 (2005).
19. Razani, B. & Lisanti, M. P. Caveolins and caveolae: molecular and functional relationships. *Exp. Cell Res.* **271**, 36–44 (2001).
20. Lu, Z., Ghosh, S., Wang, Z. & Hunter, T. Downregulation of caveolin-1 function by EGF leads to the loss of E-cadherin, increased transcriptional activity of β -catenin, and enhanced tumor cell invasion. *Cancer Cell* **4**, 499–515 (2003).
21. Di Guglielmo, G. M., Le Roy, C., Goodfellow, A. F. & Wrana, J. L. Distinct endocytic pathways regulate TGF- β receptor signalling and turnover. *Nature Cell Biol.* **5**, 410–421 (2003).
22. Elbashir, S. M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).
23. Kirchhausen, T. Clathrin. *Annu. Rev. Biochem.* **69**, 699–727 (2000).
24. Simonsen, A. *et al.* EEA1 links PI(3)K function to Rab5 regulation of endosome fusion. *Nature* **394**, 494–498 (1998).
25. Stahlhut, M. & van Deurs, B. Identification of filamin as a novel ligand for caveolin-1: evidence for the organization of caveolin-1-associated membrane domains by the actin cytoskeleton. *Mol. Biol. Cell* **11**, 325–337 (2000).
26. Tsai, B. *et al.* Gangliosides are receptors for murine polyoma virus and SV40. *EMBO J.* **22**, 4346–4355 (2003).
27. Rothman, J. E. Mechanisms of intracellular protein transport. *Nature* **372**, 55–63 (1994).
28. Pelkmans, L., Burli, T., Zerial, M. & Helenius, A. Caveolin-stabilized membrane domains as multifunctional transport and sorting devices in endocytic membrane traffic. *Cell* **118**, 767–780 (2004).
29. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
30. Zerial, M. & McBride, H. Rab proteins as membrane organizers. *Nature Rev. Mol. Cell Biol.* **2**, 107–117 (2001).
31. Pelkmans, L. & Zerial, M. Kinase-regulated quantal assemblies, kiss-and-run and recycling of caveolae. *Nature* (submitted).
32. Wary, K. K., Mariotti, A., Zurzolo, C. & Giancotti, F. G. A requirement for caveolin-1 and associated kinase Fyn in integrin signaling and anchorage-dependent cell growth. *Cell* **94**, 625–634 (1998).
33. del Pozo, M. A. *et al.* Integrins regulate Rac targeting by internalization of membrane domains. *Science* **303**, 839–842 (2004).
34. Pelkmans, L., Puntener, D. & Helenius, A. Local actin polymerization and dynamin recruitment in SV40-induced internalization of caveolae. *Science* **296**, 535–539 (2002).
35. Wellbrock, C., Karasides, M. & Marais, R. The RAF proteins take centre stage. *Nature Rev. Mol. Cell Biol.* **5**, 875–885 (2004).
36. Schlessinger, J. Cell signaling by receptor tyrosine kinases. *Cell* **103**, 211–225 (2000).
37. Davis, R. J. Signal transduction by the JNK group of MAP kinases. *Cell* **103**, 239–252 (2000).
38. Warren, G., Davoust, J. & Cockcroft, A. Recycling of transferrin receptors in A431 cells is inhibited during mitosis. *EMBO J.* **3**, 2217–2225 (1984).
39. Tuomikoski, T., Felix, M. A., Doree, M. & Gruenberg, J. Inhibition of endocytic vesicle fusion *in vitro* by the cell-cycle control protein kinase cdc2. *Nature* **342**, 942–945 (1989).
40. Minshall, R. D., Sessa, W. C., Stan, R. V., Anderson, R. G. & Malik, A. B. Caveolin regulation of endothelial function. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **285**, L1179–L1183 (2003).
41. Sharma, D. K. *et al.* Selective stimulation of caveolar endocytosis by glycosphingolipids and cholesterol. *Mol. Biol. Cell* **15**, 3114–3122 (2004).
42. Cavalli, V. *et al.* The stress-induced MAP kinase p38 regulates endocytic trafficking via the GDI:Rab5 complex. *Mol. Cell* **7**, 421–432 (2001).
43. Woodman, P. G., Mundy, D. I., Cohen, P. & Warren, G. Cell-free fusion of endocytic vesicles is regulated by phosphorylation. *J. Cell Biol.* **116**, 331–338 (1992).
44. Dubois, L., Lecourtis, M., Alexandre, C., Hirst, E. & Vincent, J. P. Regulated endocytic routing modulates wingless signaling in *Drosophila* embryos. *Cell* **105**, 613–624 (2001).
45. Xue, L. & Lucocq, J. ERK2 signalling from internalised epidermal growth factor receptor in broken A431 cells. *Cell. Signal.* **10**, 339–348 (1998).
46. Teis, D., Wunderlich, W. & Huber, L. A. Localization of the MP1-MAPK scaffold complex to endosomes is mediated by p14 and required for signal transduction. *Dev. Cell* **3**, 803–814 (2002).
47. Di Guglielmo, G. M., Baass, P. C., Ou, W. J., Posner, B. I. & Bergeron, J. J. Compartmentalization of SHC, GRB2 and mSOS, and hyperphosphorylation of Raf-1 by EGF but not insulin in liver parenchyma. *EMBO J.* **13**, 4269–4277 (1994).
48. Parton, R. G. Caveolae—from ultrastructure to molecular mechanisms. *Nature Rev. Mol. Cell Biol.* **4**, 162–167 (2003).
49. Kholodenko, B. N. Four-dimensional organization of protein kinase signaling cascades: the roles of diffusion, endocytosis and molecular motors. *J. Exp. Biol.* **206**, 2073–2082 (2003).
50. Schnell, M. J., Buonocore, L., Whitt, M. A. & Rose, J. K. The minimal conserved transcription stop-start signal promotes stable expression of a foreign gene in vesicular stomatitis virus. *J. Virol.* **70**, 2318–2323 (1996).
51. Macé, G., Miaczynska, M., Zerial, M. & Nebreda, A. R. Phosphorylation of EEA1 by p38MAP kinase regulates m opioid receptor endocytosis. *EMBO J.* (submitted).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank F. Halley and A. Kroenke for experimental assistance, J. Rose for rVSV, and G. Kochs for hybridomas expressing anti-Tag antibodies. I. Baines (Biopolis Dresden Consultants GmbH), D. Dorris (Ambion Inc.), C. Echeverri (Cenix Bioscience GmbH), R. Günther (Evotec Technologies GmbH) and M. Athellogou (Definiens AG) are acknowledged for making HT RNAi technologies and high-content, automated imaging and analysis technologies available. We thank F. Buchholz, C.-P. Heisenberg, M. Miaczynska, D. Meder, A. Schenck, A. Helenius and K. Simons for discussions and critical reading of the manuscript. L.P. would like to thank A. Helenius for continuous support. This work was supported by grants from the Max Planck Society 'RNAi interference' initiative and the Bundesministerium für Bildung und Forschung. L.P. is a Marie Curie fellow.

Author Contributions L.P. and M.Z. conceived the experimental idea. L.P. carried out the experiments with help from E.F., H.G. and M.H. Data analysis was carried out by L.P., B.H. and M.Z. L.P., E.K. and M.Z. together with I. Baines conceived and set up the HT-TDS and financed the project. L.P. and M.Z. wrote the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.Z. (zerial@mpi-cbg.de).

Photon blockade in an optical cavity with one trapped atom

K. M. Birnbaum¹, A. Boca¹, R. Miller¹, A. D. Boozer¹, T. E. Northup¹ & H. J. Kimble¹

At low temperatures, sufficiently small metallic¹ and semiconductor² devices exhibit the ‘Coulomb blockade’ effect, in which charge transport through the device occurs on an electron-by-electron basis³. For example, a single electron on a metallic island can block the flow of another electron if the charging energy of the island greatly exceeds the thermal energy. The analogous effect of ‘photon blockade’ has been proposed for the transport of light through an optical system; this involves photon–photon interactions in a nonlinear optical cavity^{4–13}. Here we report observations of photon blockade for the light transmitted by an optical cavity containing one trapped atom, in the regime of strong atom–cavity coupling¹⁴. Excitation of the atom–cavity system by a first photon blocks the transmission of a second photon, thereby converting an incident poissonian stream of photons into a sub-poissonian, anti-bunched stream. This is confirmed by measurements of the photon statistics of the transmitted field. Our observations of photon blockade represent an advance over traditional nonlinear optics and laser physics, into a regime with dynamical processes involving atoms and photons taken one-by-one.

An analogy between electron transport in mesoscopic electronic devices and photon transport through strongly coupled optical systems was originally suggested in ref. 5. These authors proposed that an effect similar to Coulomb blockade for electrons^{1–3} might be possible for photons by using photon–photon interactions in a nonlinear optical cavity⁵. In this scheme, strong dispersive interactions enabled by electromagnetically induced transparency (EIT) cause the presence of a ‘first’ photon within the cavity to block the transmission of a ‘second’ photon, leading to an ordered flow of photons in the transmitted field.

After resolution of an initial difficulty⁶, subsequent work has confirmed that such photon blockade is indeed feasible for a single intracavity atom by way of a multi-state EIT scheme^{7–9}. Photon blockade is possible in other settings, including in concert with Coulomb blockade¹⁰ and via tunnelling with localized surface plasmons¹¹. Photon blockade has also been predicted for a two-state atom coupled to a cavity mode^{4,9,12,13}. As illustrated in Fig. 1a, the underlying mechanism is the anharmonicity of the Jaynes–Cummings ladder of eigenstates^{4,15}. Resonant absorption of a photon of frequency ω_- to reach the state $|1, -\rangle$ (where $|n, +(-)\rangle$ denotes the higher- (lower-) energy eigenstate with n excitations) ‘blocks’ the absorption of a second photon at ω_- because transitions to $|2, \pm\rangle$ are detuned from resonance.

Whereas electrons interact directly via Coulomb repulsion, photon–photon interactions must be mediated by matter. Furthermore, verification of this effect requires measurements of the quantum statistics of the field; in contrast, Coulomb blockade can be inferred directly from mean transport. Scattering from a single atom in free space, for example, provides a fundamental example of photon blockade¹⁶, albeit with the fluorescent field distributed over 4π and

the flux limited by the rate of spontaneous decay γ . In contrast, cavity-mediated schemes offer the possibility of photon emission into a collimated spatial mode with high efficiency and at a rate set by the cavity decay rate κ , which can be much larger than γ . Achieving photon blockade for a single atom in a cavity requires us to operate in the regime of strong coupling, for which the frequency scale g_0 associated with reversible evolution of the atom–cavity system exceeds the dissipative rates (γ, κ) (ref. 14).

Here we report observations of photon blockade in the light transmitted by an optical cavity containing one atom strongly coupled to the cavity field. For coherent excitation at the cavity

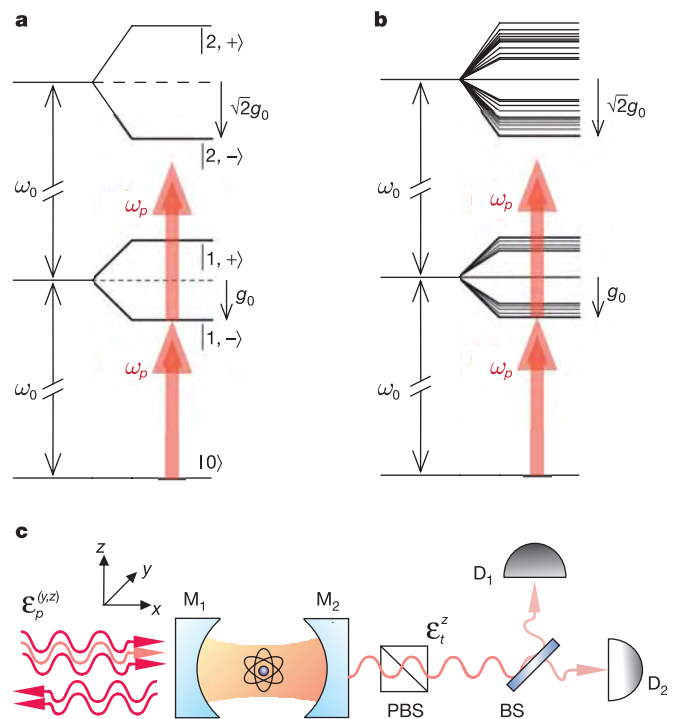


Figure 1 | The atomic level structure used for implementation of the photon blockade effect, and a simple diagram of the experiment. a, Atomic level diagram showing the lowest-energy states for a two-state atom of transition frequency ω_A coupled (with single-photon Rabi frequency g_0) to a mode of the electromagnetic field of frequency ω_C , with $\omega_A = \omega_C \equiv \omega_0$ (ref. 15). Two-photon absorption is suppressed for a probe field ω_p (arrows) tuned to excite the transition $|0\rangle \rightarrow |1, -\rangle$, $\omega_p = \omega_0 - g_0$, leading to $g^{(2)}(0) < 1$ (ref. 13). **b**, Eigenvalue structure for the $(F=4, m_F) \leftrightarrow (F'=5, m'_F)$ transition coupled to two degenerate cavity modes $l_{y,z}$, as discussed in the Supplementary Information. Two-photon absorption is likewise blocked for excitation tuned to the lowest eigenstate (arrows). **c**, Simple diagram of the experiment. BS, beam splitter.

¹Norman Bridge Laboratory of Physics 12-33, California Institute of Technology, Pasadena, California 91125, USA.

input, the photon statistics for the cavity output are investigated by measurement of the intensity correlation function $g^{(2)}(\tau)$, which demonstrates the manifestly nonclassical character of the transmitted field. Explicitly, we find $g^{(2)}(0) = (0.13 \pm 0.11) < 1$ with $g^{(2)}(0) < g^{(2)}(\tau)$, so that the output light is both subpoissonian and antibunched¹⁷. We find that $g^{(2)}(\tau)$ rises to unity at a time $\tau \approx 45$ ns, which is consistent with the lifetime $\tau_- = 2/(\gamma + \kappa) = 48$ ns for the state $|1, -\rangle$ associated with the blockade. Over longer timescales, cavity transmission exhibits modulation arising from the oscillatory motion of the atom trapped within the cavity mode. We use this modulation to make an estimate of the energy distribution for the atomic centre-of-mass motion and infer a maximum energy $E/k_B \approx 250 \mu\text{K}$, where k_B is the Boltzmann constant

The schematic of our experiment in Fig. 1c illustrates the Fabry–Perot cavity formed by mirrors (M_1, M_2) into which single optically cooled caesium atoms are loaded. Atoms are trapped within the cavity by a far-off-resonance trap (FORT), which is created by exciting a TEM₀₀ cavity mode at $\lambda_F = 935.6$ nm (ref. 18). To achieve strong coupling, we use the $6S_{1/2}, F = 4 \rightarrow 6P_{3/2}, F' = 5'$ transition of the D2 line in caesium at $\lambda_A = 852.4$ nm (subscript A refers to ‘atom’), for which the maximum rate of coherent coupling is $g_0/2\pi = 34$ MHz for $(F = 4, m_F = \pm 4) \rightarrow (F' = 5', m_F' = \pm 5)$. The transverse decay rate for the $6P_{3/2}$ atomic states is $\gamma/2\pi = 2.6$ MHz, while the cavity field decays at rate $\kappa/2\pi = 4.1$ MHz. The parameters of the cavity are further discussed in the Methods.

A variety of factors make our atom–cavity system more complex than the simple situation described by the Jaynes–Cummings eigenstates, including most significantly that (1) the cavity supports two modes $l_{y,z}$ with orthogonal linear polarizations ($\hat{y}, \hat{z}$) near $\lambda_A = 852.4$ nm as described in the Methods section, and (2) a multiplicity of Zeeman states are individually coupled to these modes for transitions between the manifolds $(F = 4, m_F) \leftrightarrow (F' = 5', m_F')$. An indication of the potential for this system to achieve photon blockade is provided in Fig. 1b, which displays the actual eigenvalue structure for the first two excited manifolds obtained by direct diagonalization of the interaction hamiltonian, as discussed in the Supplementary Information. As for the basic two-state system, excitation to the lowest-energy state in the one-excitation manifold ‘blocks’ subsequent excitation because the transitions to the two-excitation manifold are out of resonance.

To substantiate this picture quantitatively, we present in Fig. 2 theoretical results from the steady-state solution to the master equation in various situations, all for the case of coincident atomic and cavity resonances $\omega_A = \omega_{C_1} \equiv \omega_0$. (Subscripts C_1 and C_2 refer to the cavity resonances near λ_A and λ_B respectively). Beginning with the ideal setting of a two-state atom coupled to a single cavity mode, we display in Fig. 2a results for the probe transmission spectrum $T(\omega_p)$ and the intensity correlation function $g^{(2)}(0)$ of the field ϵ_t transmitted by mirror M_2 for excitation by a coherent-state probe ϵ_p of variable frequency ω_p incident upon the cavity mirror M_1 . Clearly evident in $T(\omega_p)$ are two peaks at $\omega_p = \omega_{\pm} \equiv \omega_0 \pm g_0$ associated with the vacuum-Rabi splitting for the states $|1, \pm\rangle$. At these peaks, ϵ_p is detuned for excitation $|1, \pm\rangle \rightarrow |2, \pm\rangle$, resulting in $g^{(2)}(0) < 1$ for ϵ_t . The poissonian photon statistics of the incident probe are thereby converted to subpoissonian statistics for the transmitted field by way of the photon blockade effect illustrated in Fig. 1a. For strong coupling in the weak-field limit, $g^{(2)}(0) \propto (\kappa + \gamma)^2/g_0^2$ for $\omega_p = \omega_{\pm}$ (ref. 12), hence the premium on achieving $g_0 \gg (\kappa, \gamma)$. By contrast, for $\omega_p = \omega_0 \pm g_0/\sqrt{2}$, ϵ_p is resonant with the two-photon transition $|0\rangle \rightarrow |2, \pm\rangle$, resulting in superpoissonian statistics with $g^{(2)}(0) \gg 1$. For $\omega_p = \omega_0$, there is extremely large bunching due to quantum interference between ϵ_p and the atomic polarization^{12,19}.

In Fig. 2b we examine the more complex situation relevant to our actual experiment, namely a multi-state atom coupled to two cavity modes with orthogonal polarizations $\hat{y}, \hat{z}$. Most directly related to the simple case of Fig. 2a is to excite one polarization eigenmode with the incident probe, taken here to be ϵ_p^z , and to detect the transmitted field

ϵ_t^z for this same polarization, with the transmission spectrum and intensity correlation function denoted by $T_{zz}(\omega_p)$, $g_{zz}^{(2)}(0)$, respectively. Even for the full multiplicity of states for the $F = 4 \rightarrow F' = 5'$ transition coupled to the two cavity modes $l_{y,z}$, $T_{zz}(\omega_p)$ displays a rather simple structure, now with a multiplet structure in place of the single vacuum-Rabi peak around $\omega_p \approx \omega_0 \pm g_0$. For a probe frequency tuned to the eigenvalues $\omega_p = \omega_0 \pm g_0$, $g_{zz}^{(2)}(0) \approx 0.7$, once again dropping below unity as in Fig. 2a.

An alternative scheme is to detect along $\hat{z}$, but excite along orthogonal polarization $\hat{y}$, with the respective transmission and correlation functions $T_{yz}(\omega_p)$, $g_{yz}^{(2)}(0)$ also shown in Fig. 2b. Similar to $T_{zz}(\omega_p)$, $T_{yz}(\omega_p)$ exhibits a multiplet structure in the vicinity of $\omega_p \approx \omega_0 \pm g_0$ owing to the nature of the first excited states of the atom–cavity system. At the extremal $\omega_p = \omega_0 \pm g_0$, $g_{yz}^{(2)}(0)$ reaches a value $g_{yz}^{(2)}(0) \approx 0.03$ much smaller than for either $g_{zz}^{(2)}(0)$ in Fig. 2a, or $g_{zz}^{(2)}(0)$ in Fig. 2b, for the same values of (g_0, κ, γ) . Our preliminary hypothesis is that this reduction relates to the absence of the superposed driving field ϵ_p^y with the transmitted field ϵ_t^z of orthogonal polarization $\hat{z}$ (ref. 20); photons in the mode l_z derive from emissions associated with the atomic components of atom–field eigenstates.

Tuning the probe to $\omega_p = \omega_0 \pm g_0$ has the additional benefit of reducing sensitivity to atomic position, which varies experimentally owing to atomic motion and the multiplicity of trapping sites within the cavity²¹. The atomic position affects the transmission via the position dependence of the coupling $g = g_0\psi(\mathbf{r})$, where ψ is the TEM₀₀ spatial mode at λ_{C_1} with maximum $|\psi| = 1$, and $\mathbf{r}$ is the position of the atom. $T_{yz}(\omega_p)$ is small when $|\omega_p - \omega_0| \gg g$, so atoms which have a lower-than-expected value of g will have a reduced contribution to the photon statistics.

An important step in the implementation of this strategy is our recent measurement of the vacuum-Rabi spectrum $T_{zz}(\omega_p)$ for one trapped atom²¹. In that work we obtained quantitative agreement on an atom-by-atom basis between our observations and an extension of the theoretical model used to generate the various plots in Fig. 2b. The extended model incorporates a.c.-Stark shifts from the FORT as well as cavity birefringence. This model predicts that corrections to

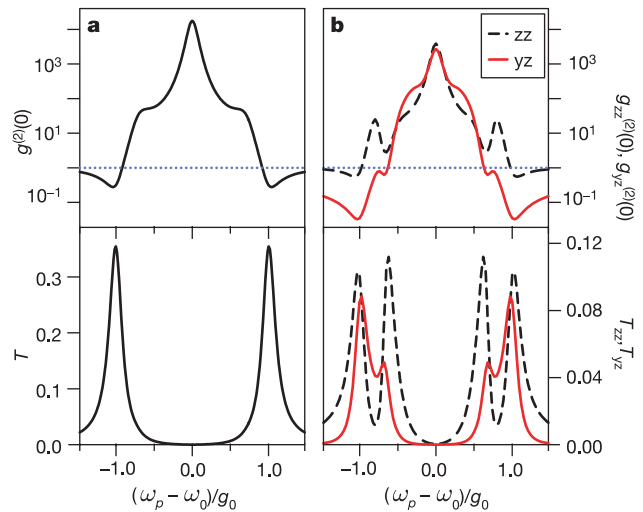


Figure 2 | Theoretical results for the transmission spectra and intensity correlation functions. **a**, $T(\omega_p)$, $g^{(2)}(0)$; **b**, $T_{zz}(\omega_p)$, $g_{zz}^{(2)}(0)$ (dashed) and $T_{yz}(\omega_p)$, $g_{yz}^{(2)}(0)$ (red) from the steady-state solution to the master equation. Included are all transitions $(F = 4, m_F) \leftrightarrow (F' = 5', m_F')$ with their respective coupling coefficients $g_0(m_F, m_F')$, as well as the two cavity modes $l_{y,z}$ here assumed to be degenerate in frequency (see Supplementary Information for further discussion). The blue dotted lines indicate poissonian statistics. Parameters are $(g_0, \kappa, \gamma)/2\pi = (33.9, 4.1, 2.6)$ MHz, and the probe strength is such that the intracavity photon number on resonance without an atom is 0.05.

$g_{yz}^{(2)}(0)$ due to these effects are small for our parameters, as discussed in the Supplementary Information.

With these capabilities, we now report measurements of $g_{yz}^{(2)}(\tau)$ for the light transmitted by a cavity containing a single trapped atom. We tune the probe ε_p to $(\omega_p - \omega_0)/2\pi = -34$ MHz, near $-g_0$, and acquire photoelectric counting statistics of the field ε_t^z by way of two avalanche photodiodes (D_1, D_2), as illustrated in Fig. 1c. From the record of these counts, we are able to determine $g_{yz}^{(2)}(\tau)$ by using the procedures discussed in ref. 22. Data are acquired for each trapped atom by cycling through probing, testing, and cooling intervals (of durations $\Delta t_{\text{probe}} = 500 \mu\text{s}$, $\Delta t_{\text{test}} = 100 \mu\text{s}$ and $\Delta t_{\text{cool}} = 1.4$ ms, respectively) using a procedure similar to that of ref. 21. The test beam is polarized along $\hat{z}$ and resonant with the cavity. A repumping beam transverse to the cavity axis and resonant with $6S_{1/2}, F=3 \rightarrow 6P_{3/2}, F'=4'$ also illuminates the atom during the probe and test intervals. This beam prevents accumulation of population in the $F=3$ ground state caused by the probe off-resonantly exciting the $F=4 \rightarrow F'=4'$ transition. All probing/cooling cycles end after an interval $\Delta t_{\text{tot}} = 0.3$ s, at which point a new

loading cycle is initiated. We select for the presence of an atom by requiring that $T_{zz}(\omega_p \approx \omega_{C_1}) \lesssim 0.35$ for the test beam. We use only those data records associated with probing intervals after which the presence of an atom was detected and for which the presence of an atom was detected in all preceding intervals. If there is no atom and the probe is tuned to be resonant with the cavity ($\omega_p = \omega_{C_1}$), then the photon number in mode l_y due to ε_p^y is 0.21 and the polarizing beam splitter at the output of the cavity (PBS in Fig. 1c) suppresses detection of this light by a factor of ~ 94 .

Figure 3 presents an example of $g_{yz}^{(2)}(\tau)$ determined from the recorded time-resolved coincidences at (D_1, D_2). In Fig. 3a, the manifestly nonclassical character of the transmitted field is clearly observed with a large reduction in $g_{yz}^{(2)}(0)$ below unity, $g_{yz}^{(2)}(0) = (0.13 \pm 0.11) < 1$, corresponding to the subpoissonian character of the transmitted field, and with $g_{yz}^{(2)}(0) < g_{yz}^{(2)}(\tau)$ as a manifestation of photon antibunching. We find that $g_{yz}^{(2)}(\tau)$ rises to unity at a time $\tau \approx 45$ ns, which is consistent with a simple estimate of $\tau_- = 2/(\gamma + \kappa) = 48$ ns based upon the lifetime for the state $|1, -\rangle$.

Although for small $|\tau|$ our observations of $g_{yz}^{(2)}(\tau)$ are in reasonable agreement with the predictions from our theoretical model, there are significant deviations on longer timescales. Modulation that is not present in the model is evident in Fig. 3b, which arises from the centre-of-mass motion of the trapped atom. In support of this assertion, Fig. 3c displays the Fourier transform $\tilde{g}(f)$ of $g_{yz}^{(2)}(\tau)$, which exhibits a narrow peak at frequency $f_0 \approx 535$ kHz just below the independently determined frequency $\nu_0 \approx 544$ kHz for harmonic motion of a trapped atom about an antinode of the FORT in the axial direction x . This modulation is analogous to that observed in ref. 23 for $g^{(2)}(\tau)$ for the light from a single ion, which arose from micro-motion of the ion in the radio-frequency trap.

Here, $U(\mathbf{r}) = U_0 \sin^2(2\pi x/\lambda_{C_2}) \exp(-2\rho^2/w_{C_2}^2)$ is the FORT potential, which gives rise to an anharmonic ladder of vibrational states with energies $\{E_m\}$. Here $m = 0$ to $m_{\text{max}} = 99$ correspond to the bound states in the axial dimension for radial coordinate $\rho \equiv \sqrt{y^2 + z^2} = 0$. The anharmonicity leads to the observed offset $f_0 < \nu_0$ due to the distribution of energies for axial motion in the FORT well. Indeed, the frequency $\nu_{\text{min}} = (E_{m_{\text{max}}} - E_{m_{\text{max}}-1})/h$ at the top of the well is approximately half that at the bottom of the well, $\nu_0 = (E_1 - E_0)/h$. By comparing the measured distribution of frequencies exhibited by $\tilde{g}(f)$ with the calculated axial frequencies $\{\nu_m\}$, we estimate that those atoms from which data was obtained are trapped in the lowest-lying axial states $m \lesssim 10$, which corresponds to a maximum energy $E/k_B \approx 250 \mu\text{K}$. This energy estimate is consistent with other measurements of $g_{yz}^{(2)}(\tau)$ that we have made, as well as the Fourier transform of the record of the transmitted intensity and the transmission spectra of ref. 21.

We have demonstrated photon blockade for the transmission of an optical cavity strongly coupled to a single trapped atom^{4-9,12,13}. The observed nonclassical photon statistics for the transmitted field result from strong nonlinear interactions at the single-photon level, in analogy with the phenomena of Coulomb blockade for electron transport¹⁻³. Extensions of our work include operation in a pulsed mode, as analysed in ref. 5, thereby realizing a source for single photons 'on demand'²². As we improve the effectiveness of our cooling procedure, we should be able to explore the dependence of $g_{yz}^{(2)}(\tau)$ on probe detuning, $\omega_p - \omega_0$, as well as to move to higher levels of excitation to increase the intracavity photon number towards unity and the output flux towards the maximum value $\approx \kappa$ for subpoissonian photons.

METHODS

Cavity and detection parameters. The physical length of the cavity used in this work is $42.2 \mu\text{m}$ and the finesse is 4.3×10^5 . The cavity length is independently stabilized such that a TEM₀₀ longitudinal mode at λ_{C_1} is resonant with the free-space atomic transition at λ_A and another TEM₀₀ mode at λ_{C_2} is resonant at λ_F . At the cavity centre $x = 0$, the mode waists $w_{C_{1,2}} = \{23.4, 24.5\} \mu\text{m}$ at $\lambda_{C_{1,2}} = \{852.4, 935.6\} \text{ nm}$.

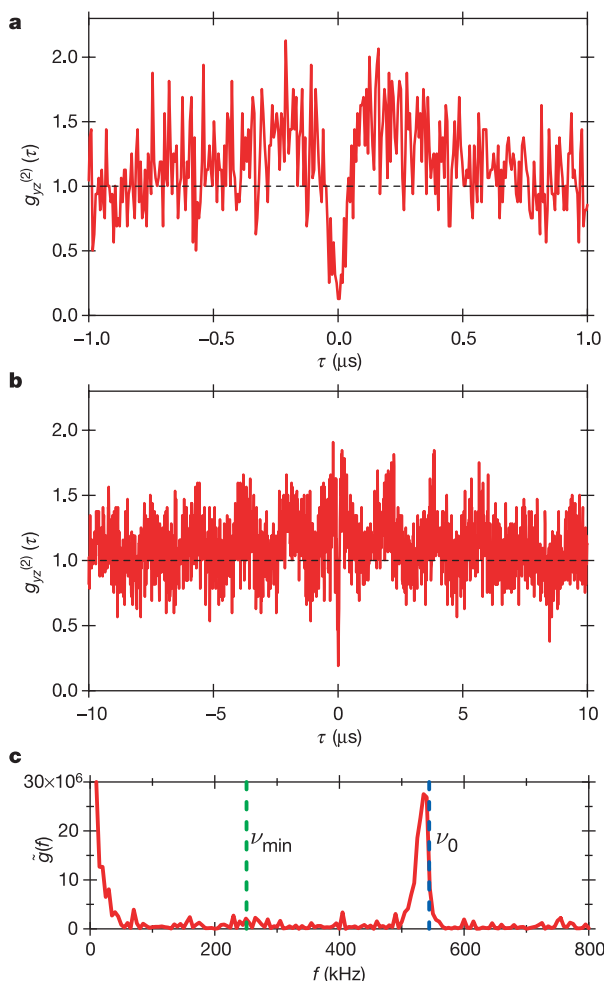


Figure 3 | Experimental measurements of the intensity correlation function $g_{yz}^{(2)}(\tau)$ for incident excitation with polarization along $\hat{y}$ and detection with orthogonal polarization $\hat{z}$. **a**, $g_{yz}^{(2)}(\tau)$ over the interval $|\tau| \leq 1.0 \mu\text{s}$ demonstrates that the transmitted field exhibits both subpoissonian photon statistics $g_{yz}^{(2)}(0) = (0.13 \pm 0.11) < 1$ and photon antibunching $g_{yz}^{(2)}(0) < g_{yz}^{(2)}(\tau)$ (ref. 17). **b**, $g_{yz}^{(2)}(\tau)$ over longer intervals $|\tau| \leq 10 \mu\text{s}$ displays a pronounced modulation due to axial motion of the trapped atom. **c**, The Fourier transform $\tilde{g}(f)$ of $g_{yz}^{(2)}(\tau)$ with the independently determined minimum and maximum frequencies ν_{min} and ν_0 for axial motion in a FORT well indicated by the dotted lines. $g_{yz}^{(2)}(\tau)$ is plotted with 6-ns resolution in **a** and with 12-ns resolution in **b**.

The TEM₀₀ longitudinal mode for the FORT is driven by a linearly polarized input field ϵ_{FORT} , resulting in nearly equal a.c. Stark shifts for Zeeman states in the $6S_{1/2}$, $F=3, 4$ manifold. At an antinode of the field, the peak value of the trapping potential for these states is $U_0/h = -43\text{MHz}$ for all our measurements. Zeeman states of the $6P_{3/2}$, $F'=5'$ manifold experience a similar trapping potential, but with a weak dependence on m_F' (ref. 18).

Stress-induced birefringence in the cavity mirrors leads to a mode splitting $\Delta\omega_{C_1}/2\pi = 4.4 \pm 0.2\text{MHz}$ of the two cavity modes $l_{y,z}$ with orthogonal linear polarizations ($\hat{y}, \hat{z}$). ϵ_{FORT} is linearly polarized and aligned along $\hat{z}$, the higher-frequency mode.

The efficiency for photon escape from the cavity, limited by losses inherent to the mirror substrates, is $\alpha_{e2} = 0.6 \pm 0.1$. The propagation efficiency from M_2 to detectors (D_1, D_2) is $\alpha_p = 0.41 \pm 0.03$, with each detector then receiving half of the photons. The avalanche photodiodes (D_1, D_2) have quantum efficiencies $\alpha_D = 0.49 \pm 0.05$.

Photon statistics. The transmission spectrum $T(\omega_p)$ is proportional to the ratio of photon flux $\langle \epsilon_t^\dagger \epsilon_t \rangle$ transmitted by M_2 to the flux $|\epsilon_p|^2$ incident upon M_1 , and normalized such that a cavity without an atom has a resonant transmission of unity, i.e. $T(\omega_p = \omega_{C_1}) = 1$. For a field with intensity operator $\hat{I}(t), g^{(2)}(\tau) \equiv \langle : \hat{I}(t)\hat{I}(t+\tau) : \rangle / \langle : \hat{I}(t) : \rangle \langle : \hat{I}(t+\tau) : \rangle$, where the colons denote time and normal ordering (ref. 17). $g_{yz}^{(2)}(\tau)$, displayed in Fig. 3a and shown with a 6-ns resolution, has been corrected for background counts due to detector dark counts and scattered light from the repumping beam. Without this correction, $g_{yz}^{(2)}(0) \approx (0.18 \pm 0.10)$ is directly derived from the recorded counts.

Received 26 March; accepted 5 May 2005.

- Fulton, T. A. & Dolan, G. J. Observation of single-electron charging effects in small tunnel junctions. *Phys. Rev. Lett.* **59**, 109–112 (1987).
- Kastner, M. A. The single-electron transistor. *Rev. Mod. Phys.* **64**, 849–858 (1992).
- Likharev, K. K. Single-electron devices and their applications. *Proc. IEEE* **87**, 606–632 (1999).
- Tian, L. & Carmichael, H. J. Quantum trajectory simulations of two-state behavior in an optical cavity containing one atom. *Phys. Rev. A* **46**, R6801 (1992).
- Imamoğlu, A., Schmidt, H., Woods, G. & Deutsch, M. Strongly interacting photons in a nonlinear cavity. *Phys. Rev. Lett.* **79**, 1467–1470 (1997).
- Grangier, P., Walls, D. F. & Gheri, K. M. Comment on “Strongly interacting photons in a nonlinear cavity”. *Phys. Rev. Lett.* **81**, 2833 (1998).
- Werner, M. J. & Imamoğlu, A. Photon-photon interactions in cavity electromagnetically induced transparency. *Phys. Rev. A* **61**, 011801 (1999).
- Rebić, S., Tan, S. M., Parkins, A. S. & Walls, D. F. Large Kerr nonlinearity with a single atom. *J. Opt. B* **1**, 490–495 (1999).
- Rebić, S., Parkins, A. S. & Tan, S. M. Polariton analysis of a four-level atom strongly coupled to a cavity mode. *Phys. Rev. A* **65**, 043806 (2002); Photon statistics of a single-atom intracavity system involving electromagnetically induced transparency. *Phys. Rev. A* **65**, 063804 (2002).
- Kim, J., Bensen, O., Kan, H. & Yamamoto, Y. A single-photon turnstile device. *Nature* **397**, 500–503 (1999).
- Smolyaninov, I. I., Zayats, A. V., Gungor, A. & Davis, C. C. Single-photon tunneling via localized surface plasmons. *Phys. Rev. Lett.* **88**, 187402 (2002).
- Brecha, R. J., Rice, P. R. & Xiao, M. N two-level atoms in a driven optical cavity: quantum dynamics of forward photon scattering for weak incident fields. *Phys. Rev. A* **59**, 2392–2417 (1999).
- Hood, C. J. *Real-time Measurement and Trapping of Single Atoms by Single Photons*. Section 6.2. PhD dissertation, California Institute of Technology (2000).
- Kimble, H. J. Strong interactions of single atoms and photons in cavity QED. *Phys. Scr.* **T76**, 127–138 (1998).
- Jaynes, E. T. & Cummings, F. W. Comparison of quantum and semiclassical radiation theories with application to the beam maser. *Proc. IEEE* **51**, 89–109 (1963).
- Kimble, H. J., Dagenais, M. & Mandel, L. Photon antibunching in resonance fluorescence. *Phys. Rev. Lett.* **39**, 691–695 (1977).
- Mandel, L. & Wolf, E. *Optical Coherence and Quantum Optics* (Cambridge Univ. Press, Cambridge, 1995).
- McKeever, J. et al. State-insensitive cooling and trapping of single atoms in an optical cavity. *Phys. Rev. Lett.* **90**, 133602 (2003).
- Carmichael, H. J., Brecha, R. J. & Rice, P. R. Quantum interference and collapse of the wavefunction in cavity QED. *Opt. Commun.* **82**, 73–79 (1991).
- Carmichael, H. J. Photon antibunching and squeezing for a single atom in a resonant cavity. *Phys. Rev. Lett.* **55**, 2790–2793 (1985).
- Boca, A. et al. Observation of the vacuum Rabi spectrum for one trapped atom. *Phys. Rev. Lett.* **93**, 233603 (2004).
- McKeever, J. et al. Deterministic generation of single photons from one atom trapped in a cavity. *Science* **303**, 1992–1994 (2004).
- Diedrich, F. & Walther, H. Nonclassical radiation of a single stored ion. *Phys. Rev. Lett.* **58**, 203–260 (1987).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We gratefully acknowledge the contributions of J. McKeever and C. J. Hood. This research is supported by the National Science Foundation, by the Caltech MURI Center for Quantum Networks, and by the Advanced Research and Development Activity (ARDA).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.J.K. (hjkimble@caltech.edu).

Doping semiconductor nanocrystals

Steven C. Erwin^{1*}, Lijun Zu^{2*}, Michael I. Haftel¹, Alexander L. Efros¹, Thomas A. Kennedy¹ & David J. Norris²

Doping—the intentional introduction of impurities into a material—is fundamental to controlling the properties of bulk semiconductors. This has stimulated similar efforts to dope semiconductor nanocrystals^{1–4}. Despite some successes^{5–11}, many of these efforts have failed, for reasons that remain unclear. For example, Mn can be incorporated into nanocrystals of CdS and ZnSe (refs 7–9), but not into CdSe (ref. 12)—despite comparable bulk solubilities of near 50 per cent. These difficulties, which have hindered development of new nanocrystalline materials^{13–15}, are often attributed to ‘self-purification’, an allegedly intrinsic mechanism whereby impurities are expelled. Here we show instead that the underlying mechanism that controls doping is the initial adsorption of impurities on the nanocrystal surface during growth. We find that adsorption—and therefore doping efficiency—is determined by three main factors: surface morphology, nanocrystal shape, and surfactants in the growth solution. Calculated Mn adsorption energies and equilibrium shapes for several nanocrystals lead to specific doping predictions. These are confirmed by measuring how the Mn concentration in ZnSe varies with nanocrystal size and shape. Finally, we use our predictions to incorporate Mn into previously undopable CdSe nanocrystals. This success establishes that earlier difficulties with doping are not intrinsic, and suggests that a variety of doped nanocrystals—for applications from solar cells¹⁶ to spintronics¹⁷—can be anticipated.

When a macroscopic semiconductor crystal is grown under conditions of thermal equilibrium, impurity atoms can be incorporated up to their solid solubility limit—as much as 50% or more for Mn in II–VI semiconductors¹⁸. This thermodynamic limit is completely determined by the Gibbs free energy (approximately equal to the impurity formation energy) and the growth temperature. For semiconductor nanocrystals—which are typically grown using colloidal synthesis—the impurity concentrations attained in experiments are much lower than expected from this limit^{7,8}, and for some materials are even zero¹². The likely reason is that thermal equilibrium, which requires facile diffusion, may be far from realized. Indeed, at the temperatures used in colloidal growth, typically around 300 °C, diffusion of Mn in II–VI semiconductors is negligible; for example, the diffusion length of Mn in CdTe after 1 h at this temperature is only 1–3 Å (ref. 19). This suggests that thermal equilibrium is an inappropriate starting point for describing doping in nanocrystals.

We propose instead a simple model of doping based on kinetics. Specifically, we hypothesize that impurities are incorporated into a nanocrystal only if they can bind to its surface for a residence time comparable to the reciprocal growth rate. In this model, impurity diffusion within the nanocrystal plays no role, even though diffusion (of both impurities and host atoms) on the nanocrystal surface may occur. To explore the consequences of the model, we make two reasonable assumptions: first, that the nanocrystal surface consists of

well-defined facets (as observed experimentally^{20,21}), and second, that the binding energy of an impurity adsorbed on a given facet determines its residence time. Figure 1 shows theoretical binding energies, calculated using density-functional theory^{22,23} (see Methods), for Mn adsorbed on the three most important facets of six representative II–VI and IV–VI semiconductors. Of particular interest are the binding energies on the (001) surfaces of crystals with the zinc-blende structure. These energies are strikingly larger, by factors of 2–10, than on the other two zinc-blende orientations, or on any facet of crystals with the wurtzite or rock-salt structures. The origin of this difference is the unique morphology of (001) surfaces of zinc-blende (and diamond) crystals, which typically consists of various arrangements of anion dimers. These dimers provide very stable binding sites that are absent from the (110) and (111) surfaces of zinc-blende crystals, and from all surfaces of wurtzite and rock-salt crystals. The resulting strong binding indicates that the (001) facets of zinc-blende nanocrystals play a special role in the doping process.

Indeed, a survey of the experimental literature reveals that all of the nanocrystals that have been successfully doped with individual Mn atoms (ZnS, ZnSe, CdS) exhibit the zinc-blende crystal structure^{5–9}. In contrast, those that have the wurtzite (CdSe) or rock-salt structure (PbS, PbSe) have either not been successfully doped, or have required polychalcogenide precursors to incorporate Mn in cluster form as a prebonded Se–Mn complex (a technique beyond the scope of our model)^{10,12,24}. The correspondence between these earlier results and our adsorption findings strongly suggests a new view of doping: that individual Mn impurities are most easily incorporated in nanocrystals with the zinc-blende crystal structure, by adsorbing on their (001) facets.

To explore in detail the consequences of this view, we begin by developing a model describing how doping efficiency varies over the range of possible nanocrystal shapes. For nanocrystals having the zinc-blende structure and more than one facet orientation, we assume that only the regions bounded by (001) facets can be doped. Hence, the dopable fraction, f , of the nanocrystal volume is given by the relative surface area of its (001) facets. For nanocrystals grown out of equilibrium, f can only be determined directly from knowledge of the nanocrystal shape. However, for a nanocrystal that assumes its equilibrium crystal shape, f can be determined analytically from the surface energies, E , of its possible facets. As Fig. 2 shows, (001) facets will be present (that is, $f > 0$) only if these energies satisfy $E_{110}/E_{100} > 1/\sqrt{2}$ and $E_{111}/E_{100} > 1/\sqrt{3}$. The limiting case of a nanocrystal consisting entirely of (001) facets (that is, $f = 1$) occurs when $E_{110}/E_{100} = \sqrt{2}$ and $E_{111}/E_{100} = \sqrt{3}$. For intermediate surface-energy ratios, the stable facets and relative areas shown in Fig. 2 were determined by finding the crystal shape that minimizes the total surface energy within the constraint of fixed volume.

To predict the value of f for a specific material, we must calculate absolute surface energies for all the facet orientations considered. For

¹Naval Research Laboratory, Washington DC 20375, USA. ²Department of Chemical Engineering & Materials Science, University of Minnesota, Minneapolis, Minnesota 55455, USA.

*These authors contributed equally to this work.

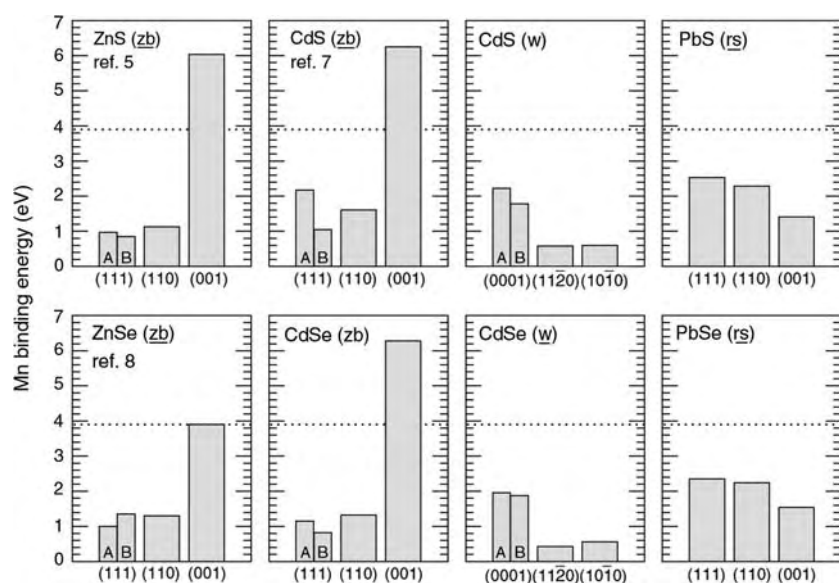


Figure 1 | Theoretical binding energies for individual Mn adsorbates on various semiconductor surfaces. The crystal structure of these semiconductors is either zinc blende (zb), wurtzite (w) or rock salt (rs); the most commonly occurring structure in nanocrystal form is underlined. Binding energies were calculated for reconstructed surfaces (see Methods) of the three most common crystallographic orientations; 'A' and 'B' denote inequivalent terminations of surfaces having the same orientation. For comparison, the binding energy per atom of bulk crystalline Mn is shown by the dotted line. Nanocrystals that have been successfully doped with individual Mn are indicated by appropriate citations.

a compound semiconductor, it is important to note that the surface energy is not uniquely defined, but rather depends on the relative concentration of its constituents through their chemical potentials. This raises the interesting possibility of controlling the nanocrystal shape—and thereby its doping efficiency—by varying these concentrations. For the example of ZnSe, by assuming specific surface morphologies (for simplicity, we have used reconstructions based on bulk crystal surfaces; see Methods) we can predict the allowed range of expected equilibrium crystal shapes at zero temperature as the chemical potentials are varied between the Zn-rich and Se-rich limits. As shown in Fig. 2, the predicted range of f is between zero (Zn-rich) and 0.3 (Se-rich). Of course, at finite temperatures, this geometrical construction must be modified to account for thermal rounding and related effects²⁵, but in principle f can still be sensibly defined.

To test these predictions experimentally, we prepared Mn-doped ZnSe nanocrystals with diameters in the range 25–50 Å (ref. 8). After initial nucleation at 310 °C, nanocrystal growth proceeded at temperatures in the range 260–300 °C, using two different Mn:Zn ratios in solution (atomic ratios, 0.05 and 0.025). For each, several different Se:Zn ratios were used (in the range 0.5–4.0) to investigate how doping efficiency varied with Se chemical potential.

In large nanocrystals, Mn incorporation can be measured directly, using inductively coupled plasma atomic emission spectroscopy (ICP). For nanocrystals of diameter 50 Å grown with Se:Zn = 1, we measured the concentrations of incorporated Mn to be 0.23% (for Mn:Zn = 0.025) and 0.45% (for Mn:Zn = 0.05). This linear scaling of Mn concentration in the nanocrystal with Mn:Zn ratio in solution is consistent with an adsorption process characterized by a very high sticking probability, in agreement with the large theoretical Mn binding energies (4–6 eV) for zinc-blende (001) facets.

Both of these Mn:Zn ratios lead to the same doping efficiency, $f = 0.09$, well within the predicted range of allowed doping fraction for ZnSe. For nanocrystals grown using enhanced anion/cation ratios (Se:Zn = 3–4) the measured Mn concentrations were indeed higher, 0.30% (for Mn:Zn = 0.025) and 0.60% (for Mn:Zn = 0.05), a doping efficiency of $f = 0.12$. We attribute this 30% increase in doping efficiency to changes in the equilibrium crystal shape. In other words, by increasing the Se:Zn ratio, the Se chemical potential was raised, which in turn reduced the surface energy of the Se-rich (001) facets, increasing their relative surface area.

In small nanocrystals, photoluminescence spectroscopy provides a more sensitive probe of Mn impurity concentration than ICP. Optically excited Mn-doped ZnSe nanocrystals exhibit two photoluminescence lines, as shown in Fig. 3a. One is due to an internal Mn^{2+} transition, and the other from exciton recombination at the ZnSe band edge. The average number of Mn atoms in the nanocrystal, $\bar{N}$, can be determined from the intensity ratio of these emission lines, $I_{\text{Mn}}/I_{\text{ZnSe}}$, while the size of the nanocrystal can be determined from the spectral shift (relative to bulk) of the ZnSe emission line. Hence, we can obtain detailed information about the distribution of Mn within the nanocrystals by analysing how $I_{\text{Mn}}/I_{\text{ZnSe}}$ varies with $\bar{N}$.

The intensity ratios in Fig. 3b demonstrate that Mn incorporation rises monotonically with increasing Se:Zn ratio, as predicted, for a wide range of nanocrystals. Interestingly, $I_{\text{Mn}}/I_{\text{ZnSe}}$ falls rapidly to zero for nanocrystals smaller than ~20 Å, suggesting a central core that resists doping. This is not unexpected, since the assumption of a faceted equilibrium shape must eventually break down for small nanocrystals. Indeed, in this size regime many semiconductors form non-crystalline cage-like clusters^{26,27}, typically with highly stable surfaces that suppress impurity adsorption. Accordingly, we model the distribution of Mn in the nanocrystal as shown in Fig. 3c and d.

The theoretical binding energies of Fig. 1 also suggest that CdSe nanocrystals (which normally have the wurtzite structure) would be easier to dope if they could be grown in the zinc-blende structure. To test this prediction, we grew core/shell nanocrystals in which a shell

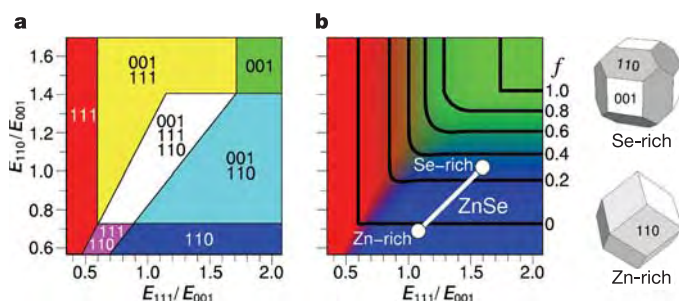


Figure 2 | Equilibrium crystal shape for cubic systems, as determined by the ratios of their surface energies. **a**, Phase diagram specifying which facets are present in the equilibrium crystal shape. **b**, Relative surface area of (111), (001) and (110) facets (red, green and blue shading, respectively) in the equilibrium crystal shape. Contours of constant relative (001) facet area are shown; this quantity gives the relative doping fraction, f , for nanocrystals having the zinc-blende structure. Numbers on the right label the contours. For ZnSe, calculated surface energies lead to the indicated range of equilibrium crystal shapes; the extremal crystal shapes in the Se- and Zn-rich limits are shown at right.

of CdSe was deposited on a ZnSe core²⁸, as shown in Fig. 4a. Because ZnSe has the zinc-blende structure, the CdSe shell can be expected to adopt this structure. Under certain growth conditions this indeed occurs, as confirmed by the X-ray diffraction data in Fig. 4b. Using these conditions, we grew CdSe shells on undoped ZnSe cores, using a growth solution containing Mn, then used electron paramagnetic resonance (EPR) to determine the location of the Mn (ref. 8). The result, shown in Fig. 4c along with control spectra, shows a six-line spectrum with a hyperfine splitting expected for Mn at lattice sites in cubic CdSe (ref. 29), consistent with the prediction that Mn is more easily incorporated into zinc-blende CdSe.

Finally, we address the role of surfactants in nanocrystal doping. The Mn binding energies shown in Fig. 1 were calculated for clean surfaces. These are relevant energies because during growth, the surfactant molecules detach from the nanocrystal, allowing adsorption onto the clean surface underneath. But the surfactants themselves can also bind Mn, competing with surface adsorption and quenching the doping. This may become important in cases where the Mn–substrate binding is only moderately strong. For example, the binding energy of Mn on the (0001) surface of wurtzite CdSe is two times smaller than on ZnSe(001). This suggests that previous failures to dope CdSe may have resulted not only from intrinsic

properties of CdSe, but also from parasitic binding of Mn by strong surfactants (such as phosphonic acids³⁰). It also raises the possibility that the incorporation of Mn into previously undopable wurtzite CdSe nanocrystals may be feasible, although this is likely to be more difficult than in ZnSe.

To examine this possibility, we grew CdSe nanocrystals in the presence of Mn (Mn:Cd = 0.05) under the same surfactant conditions used for ZnSe. According to X-ray diffraction (Fig. 4d) and EPR (Fig. 4e) data, the resulting CdSe nanocrystals have the wurtzite structure and exhibit a hyperfine splitting expected for Mn at hexagonal lattice sites in CdSe (ref. 29). Although this clearly indicates successful doping, ICP measurements also show that the Mn incorporated was only 0.14%, three times less than under identical conditions for ZnSe. We attribute this reduction to the smaller Mn adsorption binding energy, a nanocrystal shape with less adsorbing surface area, or both.

Although we have limited our discussion to II–VI and IV–VI nanocrystals and Mn impurities, the most general formulation of our model—that doping is controlled by impurity adsorption on the nanocrystal surface—provides a framework appropriate for other semiconductor families and impurities as well. For example, III–V

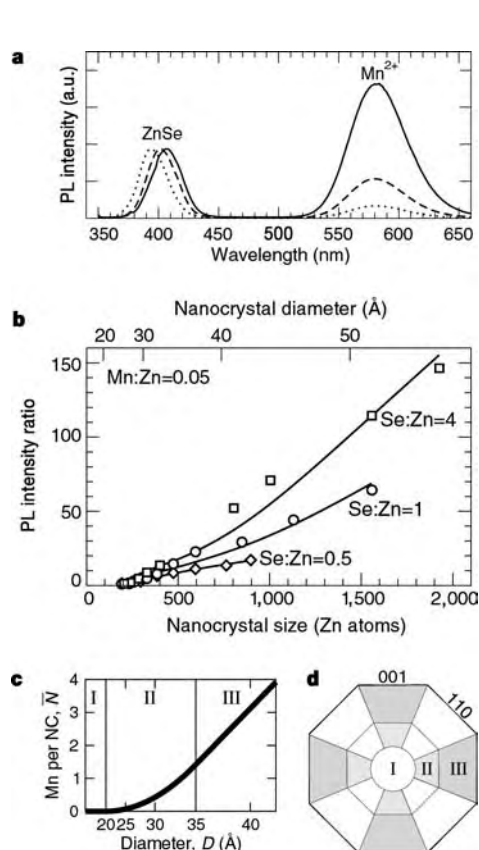


Figure 3 | Photoluminescence data and theoretical doping model for ZnSe nanocrystals doped with Mn. **a**, Typical photoluminescence (PL) spectra for three different nanocrystal sizes⁸. **b**, Dependence of intensity ratios on nanocrystal size for different Se:Zn ratios. Symbols are experimental data; curves are fits to our optical excitation model (see Methods in Supplementary Information). **c**, Model describing radial distribution of Mn incorporated in a nanocrystal: for diameter $D < 20$ Å (region I), no Mn is incorporated; for $D > 35$ Å (region III), every site in the fraction f of the nanocrystal volume is doped with probability x , taken equal to the Mn:Zn ratio in the growth solution; for the transitional region 20 Å $< D < 35$ Å (region II), a smooth interpolation is assumed. For the most favourable doping conditions used, our largest nanocrystals contain on average 12 Mn atoms. NC, nanocrystal. **d**, Cross-sectional view of a schematic nanocrystal, with dopable regions shaded (for clarity, $f = 0.5$ is used in this view).

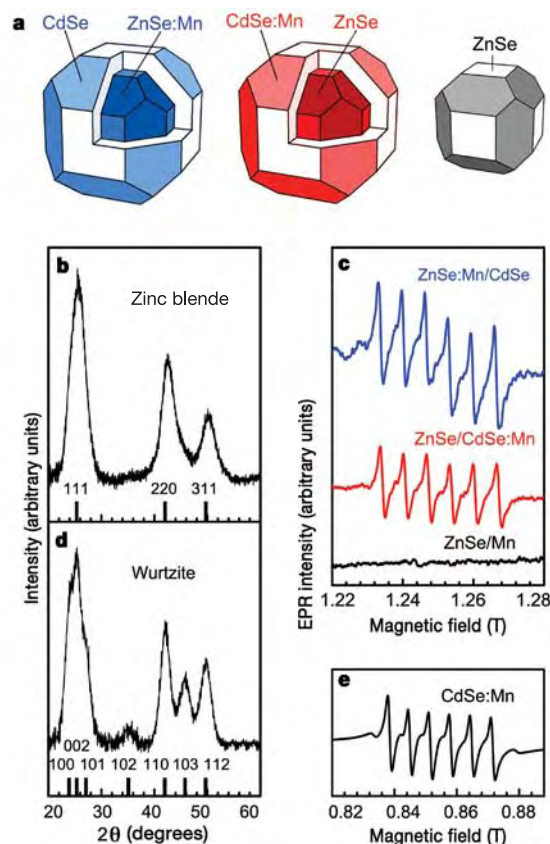


Figure 4 | Mn doping of zinc-blende and wurtzite CdSe nanocrystals. **a**, ZnSe/CdSe:Mn core/shell nanocrystals (red) and two control samples, ZnSe:Mn/CdSe core/shell nanocrystals (blue), and ZnSe nanocrystals heated in the presence of Mn (grey). The latter were grown to rule out the possibility that Mn entered the ZnSe core during the CdSe shell growth. In all three samples, the ZnSe core diameter was 36 Å. **b**, X-ray diffraction pattern for ZnSe/CdSe:Mn core/shell nanocrystals. The expected peak positions for zinc-blende CdSe are indicated. **c**, 35-GHz EPR spectra for the three types of nanocrystals illustrated in **a**. The hyperfine splittings are 65.6 G (blue) and 65.9 G (red). Mn was not detected in the second control sample (black). **d**, X-ray diffraction pattern for CdSe:Mn nanocrystals. The expected peak positions for wurtzite CdSe are indicated. **e**, 24-GHz EPR spectra for the wurtzite CdSe:Mn nanocrystals. The hyperfine splitting is 66.1 G.

and group-IV nanocrystals may, because of their increased covalency, exhibit a variety of surface morphologies quite different from their bulk reconstructions. Nevertheless, those morphologies—if known—can still provide detailed insight into doping. Indeed, for any nanocrystal with well-defined facets, differential adsorption offers a new route to the control of doping. In this way, recent experimental advances in controlling the shape³⁰ and crystal structure of nanocrystals may also become extremely important in optimizing the efficiency with which the nanocrystals can be doped.

METHODS

Theoretical. Surface energies and Mn-adsorbate binding energies were calculated within density-functional theory in the generalized-gradient approximation²³, using the projector-augmented-wave method as implemented in VASP²². Surfaces were represented by eight-layer (or thicker) slabs with the theoretical equilibrium lattice constant, reconstructed on one side and appropriately passivated on the other. Surface reconstructions from the literature were used when available; otherwise, separate studies were performed to establish the lowest-energy reconstruction from a pool of likely candidates (see discussion in Supplementary Information). Full atomic relaxation was performed for all but the bottom semiconductor layer, using 4 × 4 sampling of the surface Brillouin zone. For the (spin-polarized) adsorbate calculations, 2 × 2 supercells were used to represent isolated Mn adsorbates; within this cell, binding energies (using 2 × 2 sampling) at different adsorption sites were compared and the largest value reported. For surfaces that have more than one stable reconstruction within the allowed range of chemical potential, Mn binding energies are reported for the anion-rich reconstruction. Absolute surface energies for ZnSe(001) and (110) were computed using symmetric slabs. For the inequivalent (111)A and B orientations this is not possible, so we approximated the (111)A and B surface energies by their average. (See the discussion in Supplementary Information for details about the ZnSe reconstructions.) The Se- and Zn-rich limits of the chemical potential were defined by the Se (A8) phase of Se and the hexagonal close-packed phase of Zn, respectively.

Experimental. Mn-doped ZnSe nanocrystals were synthesized as previously reported⁸. To change the ratio of Se to Zn in the reaction, the number of moles of each was adjusted while constraining the total number of moles. The amount of Mn added was relative to the Zn concentration. To obtain the largest nanocrystals, additional precursors (in the same ratios) were sometimes added to avoid saturation in the growth. CdSe shells were deposited on ZnSe cores by modifying the procedure of ref. 28. First, ZnSe cores were prepared by injecting 4 ml of triethylphosphine (TOP), 1 ml of 1-M Se in TOP, and 82 μl of diethylzinc into 10 ml of distilled and degassed hexadecylamine (HDA) at 310 °C. (For Mn-doped ZnSe cores, 1.0 ml of 0.04-M dimethylmanganese in a tetrahydrofuran/toluene mixture⁸ was included in the injection.) The mixture was then heated at 260–300 °C until the first absorption peak was at 400 nm. To coat with CdSe, 3 ml of this solution was combined with 6 ml of distilled and degassed HDA and heated to 270 °C, followed by dropwise addition (over 30 min) of a mixture of 45 μl of dimethylcadmium, 2.4 ml of 1-M Se in TOP, and 4 ml of TOP. (For Mn-doped CdSe shells, 0.75 ml of 0.04-M dimethylmanganese was included in the addition. For the second control sample, the addition included only 0.75 ml of 0.04-M dimethylmanganese and 2 ml of TOP.) Afterwards, the solution was stirred at 150 °C for 24 h. Mn-doped CdSe nanocrystals were obtained by injecting 4 ml of TOP, 0.8 ml of 1-M Se in TOP, 60 μl of dimethylcadmium, and 1.0 ml of freshly prepared 0.04-M dimethylmanganese into 12 ml of distilled and degassed HDA at 300 °C. The mixture was then heated at 270 °C until the first absorption peak reached 600 nm. Owing to the weak binding of HDA, the CdSe nanocrystals easily formed aggregates.

Before characterization with ICP and EPR, all samples were precipitated and repeatedly washed with pyridine^{8,12} to remove excess Mn at the particle surface, as confirmed by EPR. ICP was performed on a Perkin-Elmer Optima 3000DV after dissolving the nanocrystals in aqua regia. Photoluminescence was collected with a Spex Fluorolog-2 spectrofluorometer. X-ray diffraction was measured with a Panalytical X'Pert diffractometer using Cu K radiation at 45 kV and 40 mA after the nanocrystals had been precipitated three times with methanol, dispersed in 10% octane in hexane, and then spin-coated onto silicon wafers. For EPR, the nanocrystals were cast in poly(lauryl methacrylate) films or collected as a powder. Room temperature spectra were measured at 35 and 24 GHz.

Received 14 January; accepted 10 May 2005.

- Ekimov, A. I. & Onushchenko, A. A. Quantum size effect in 3-dimensional microscopic semiconductor crystals. *JETP Lett.* **34**, 345–349 (1981).

- Efros, A. L. & Efros, A. L. Interband absorption of light in a semiconductor sphere. *Sov. Phys. Semicond.* **16**, 772–775 (1982).
- Brus, L. E. A simple model for the ionization potential, electron affinity, and aqueous redox potentials of small semiconductor crystallites. *J. Chem. Phys.* **79**, 5566–5571 (1983).
- Alivisatos, A. P. Semiconductor clusters, nanocrystals, and quantum dots. *Science* **271**, 933–937 (1996).
- Wang, Y., Herron, N., Moller, K. & Bein, T. 3-dimensionally confined diluted magnetic semiconductor clusters: Zn_{1-x}Mn_xS. *Solid State Commun.* **77**, 33–38 (1991).
- Bhargava, R. N., Gallagher, D., Hong, X. & Nurmikko, A. Optical properties of manganese-doped nanocrystals of ZnS. *Phys. Rev. Lett.* **72**, 416–419 (1994).
- Levy, L., Hochepiet, J. F. & Pileni, M. P. Control of the size and composition of three dimensionally diluted magnetic semiconductor clusters. *J. Phys. Chem.* **100**, 18322–18326 (1996).
- Norris, D. J., Yao, N., Charnock, F. T. & Kennedy, T. A. High-quality manganese-doped ZnSe nanocrystals. *Nano Lett.* **1**, 3–7 (2001).
- Suyver, J. F., Wuister, S. F., Kelly, J. J. & Meijerink, A. Luminescence of nanocrystalline ZnSe:Mn²⁺. *Phys. Chem. Chem. Phys.* **2**, 5445–5448 (2000).
- Hanif, K. M., Meulenberg, R. W. & Strouse, G. F. Magnetic ordering in doped Cd_{1-x}Co_xSe diluted magnetic quantum dots. *J. Am. Chem. Soc.* **124**, 11495–11502 (2002).
- Stowell, C. A., Wiacek, R. J., Saunders, A. E. & Korgel, B. A. Synthesis and characterization of dilute magnetic semiconductor manganese-doped indium arsenide nanocrystals. *Nano Lett.* **3**, 1441–1447 (2003).
- Mikulec, F. V. et al. Organometallic synthesis and spectroscopic characterization of manganese-doped CdSe nanocrystals. *J. Am. Chem. Soc.* **122**, 2532–2540 (2000).
- Shim, M. & Guyot-Sionnest, P. n-type colloidal semiconductor nanocrystals. *Nature* **407**, 981–983 (2000).
- Hoffman, D. M. et al. Giant internal magnetic fields in Mn doped nanocrystal quantum dots. *Solid State Commun.* **114**, 547–550 (2000).
- Efros, A. L., Rashba, E. I. & Rosen, M. Paramagnetic ion-doped nanocrystal as a voltage-controlled spin filter. *Phys. Rev. Lett.* **87**, 206601 (2001).
- Huynh, W. U., Dittmer, J. J. & Alivisatos, A. P. Hybrid nanorod-polymer solar cells. *Science* **295**, 2425–2427 (2002).
- Wolf, S. A. et al. Spintronics: a spin-based electronics vision for the future. *Science* **294**, 1488–1495 (2001).
- Hwang, I. S., Kim, H. D., Kim, J. E., Park, H. Y. & Lim, H. Solid solubilities of magnetic ions in diluted magnetic semiconductors grown under equilibrium conditions. *Phys. Rev. B* **50**, 8849–8852 (1994).
- Jamil, N. Y. & Shaw, D. The diffusion of Mn in CdTe. *Semicond. Sci. Technol.* **10**, 952–958 (1995).
- Shiang, J. J., Kadavanich, A. V., Grubbs, R. K. & Alivisatos, A. P. Symmetry of annealed wurtzite CdSe nanocrystals — assignment to the C_{3v} point group. *J. Phys. Chem.* **99**, 17417–17422 (1995).
- Peng, X. G., Wickham, J. & Alivisatos, A. P. Kinetics of II–VI and III–V colloidal semiconductor nanocrystal growth: ‘Focusing’ of size distributions. *J. Am. Chem. Soc.* **120**, 5343–5344 (1998).
- Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
- Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
- Ji, T. H., Jian, W. B. & Fang, J. Y. The first synthesis of Pb_{1-x}Mn_xSe nanocrystals. *J. Am. Chem. Soc.* **125**, 8448–8449 (2003).
- Wortis, M. *Chemistry and Physics of Solid Surfaces VII* Ch. 13 (Springer, Berlin, 1988).
- Kasuya, A. et al. Ultra-stable nanoparticles of CdSe revealed from mass spectrometry. *Nature Mater.* **3**, 99–102 (2004).
- Kaxiras, E. Effect of surface reconstruction on stability and reactivity of Si clusters. *Phys. Rev. Lett.* **64**, 551–554 (1990).
- Balet, L. P., Ivanov, S. A., Piryatinski, A., Achermann, M. & Klimov, V. I. Inverted core/shell nanocrystals continuously tunable between type-I and type-II localization regimes. *Nano Lett.* **4**, 1485–1488 (2004).
- Koh, A. K. & Miller, D. J. The systematic variation of the EPR parameters of manganese in II–VI semiconductors. *Solid State Commun.* **60**, 217–222 (1986).
- Peng, X. G. et al. Shape control of CdSe nanocrystals. *Nature* **404**, 59–61 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the US Office of Naval Research, the NSF-MRSEC at the University of Minnesota, and NSF-CTS. Computations were performed at the Department of Defense Major Shared Resource Center at ASC. We thank Y. Nesmelov, P. Hasjim and R. Weber for experimental assistance.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.C.E. (Steven.Erwin@nrl.navy.mil) or D.J.N. (dnorris@umn.edu).

Breaking of Henry's law for noble gas and CO₂ solubility in silicate melt under pressure

Philippe Sarda^{1†} & Bertrand Guillot²

Degassing of the Earth is still poorly understood, as is the large scatter in He/Ar ratios observed in mid-ocean ridge basalts. A possible explanation for such observations is that vesiculation occurs at great depths with noble-gas solubilities different from those measured at 1 bar (ref. 1). Here we develop a hard-sphere model for noble-gas solubility and find that, owing to melt compaction, solubility may decrease by several orders of magnitude when pressure increases, an effect subtly overbalanced by the compression of the fluid phase. Our results satisfactorily explain recent experimental data on argon solubility in silicate melts, where argon concentration increases almost linearly with pressure, then levels off at pressures of 50–100 kbar (refs 2–5). We also model vesiculation during magma ascent at ridges and find that noble-gas partitioning between melt and CO₂ vesicles at depth differs significantly from that at low pressure. Starting at 10 kbar (~35 km depth), several stages of vesiculation occur followed by vesicle loss, which explains the broad variability of He–Ar concentration data in mid-ocean ridge basalts. ‘Popping rocks’, exceptional samples with high vesicularity, may represent fully vesiculated ridge magma, whereas common samples would simply have lost such vesicles.

The solubility of a gas in a liquid is described by the equality of its chemical potential in the two phases at equilibrium. For a rare gas fluid in contact with a silicate melt at fixed temperature T and pressure P , the above equality leads to the relationship:

$$\rho_m/\rho_g = e^{-\beta(\mu_{\text{ex},m} - \mu_{\text{ex},g})} = \gamma_m/\gamma_g \quad (1)$$

where ρ_m and ρ_g are the number densities (particles per unit volume) of the solute gas in the melt and gas phase, respectively, $\beta = 1/k_B T$ is the reciprocal of temperature with k_B the Boltzmann constant and where $\mu_{\text{ex},m}$ and $\mu_{\text{ex},g}$ are the excess chemical potentials (with respect to the ideal gas) of the solute in the respective two phases. Coefficient $\gamma_i = e^{-\beta\mu_{\text{ex},i}}$ (where i indicates either m or g) is called the solubility parameter of the rare gas in the corresponding phase. Experimentally, it is the mole fraction (often called concentration), $X = \rho_m/(\rho_m + \rho_s)$, of noble gas in the solvent melt, the number density of which is ρ_s , that is measured (or, alternatively, the weight fraction). Using equation (1), we can say:

$$X = L/(1 + L) \quad (2)$$

where $L = \rho_g \gamma_m / \rho_s \gamma_g$. The above equation is general. Only at low pressure can gas be considered as nearly ideal (that is $\gamma_g \approx 1$ and $P_g \approx \rho_g k_B T$), and the mole fraction be approximated by the expression:

$$X = P_g S \quad (3)$$

where $S = \gamma_m / \rho_s k_B T$ is the solubility constant (the inverse of the Henry constant) expressed in bar⁻¹. When pressure increases, the imperfection of gas is taken into account by γ_g and ρ_g in equation

(2). However, expression (3) is often considered valid provided that fugacity, $f = P_g e^{+\beta\mu_{\text{ex},g}}$, is substituted for pressure, P_g . This procedure is approximate and somewhat misleading because the relation between P_g and ρ_g is then not correctly accounted for. In the kilobar range and above, neither the pressure dependence of γ_m and γ_g nor the proper variation of ρ_g and ρ_s can be neglected.

A heuristic method is to evaluate X in the framework of the hard-sphere fluid, the reference model in liquid state theory⁶, where atoms (or molecules) are described by hard spheres of diameter d , and the equation of state is given by the Carnahan–Starling equation⁷. The latter reproduces the compressibility factor ($P_g/\rho_g k_B T$) very well and the excess chemical potential ($\mu_{\text{ex},g}$) of simple gases in the supercritical region, provided a temperature-dependent hard-sphere diameter is used⁸ (for details see the Supplementary Information). More surprising is that the compressibility of silicate melts can also be conveniently described by the hard-sphere fluid, where, to account for the cohesion energy, the equation of state is scaled to the melt density at a pressure of 1 bar and at the temperature of investigation (a density calculated from literature data⁹). After fitting the compressibility data of various silicate melts, the hard-sphere diameter is found to decrease from silicic (highly polymerized) to ultrabasic (weakly polymerized) melts. Consequently, their propensity to accommodate a given noble gas (in voids between spheres) should decrease in the same order.

The solubility parameter γ_m is next evaluated from statistical thermodynamics according to the scaled particle theory developed for hard spheres¹⁰, describing an infinitely diluted solution of noble gas in the liquid silicate, where γ_m is the probability of inserting the solute particle into the density fluctuations of the pure solvent. The excess chemical potential (see equation (1)) is the sum of two contributions: the entropy of cavity formation and the solvation energy after insertion. The scaled particle theory yields the entropy term (strongly pressure dependent) while the solvation energy (weakly pressure dependent) is accounted for through the electronic polarization induced by the ionic melt (see Supplementary Information). This ansatz has shown its ability to calculate noble-gas solubilities in computer-simulated silica¹¹.

We first evaluate the solubility parameters, γ_m and γ_g , of pure He, Ne, Ar and Xe in a tholeiitic melt, using the published solubility constants¹² as constraints at 1 bar. As shown in Fig. 1, in the kilobar range and beyond, compaction of melt and the rare gas fluid cannot be neglected: γ_m and γ_g decrease by several orders of magnitude between 1 and 100 kbar, the larger the rare gas the stronger the decrease. More surprising is the behaviour of the ratio γ_m/γ_g , which determines concentration X (see equation (2), $X \approx L$). This ratio increases with pressure and tends to saturate for He and Ne above 100 kbar, while for Ar and Xe it goes through a maximum around 70–80 kbar, then drops off at higher pressures. The increase of

¹Groupe Géochimie des Gaz Rares, Département des Sciences de la Terre, Université Paris Sud, UMR CNRS 8148 (IDES), 91405 Orsay Cedex, France. ²Laboratoire de Physique Théorique de la Matière Condensée, Université Pierre et Marie Curie, UMR CNRS 7600, 4 place Jussieu, 75252 Paris Cedex 05, France. †Present address: Laboratoire de Sciences de la Terre, Ecole Normale Supérieure de Lyon, UMR CNRS 5570, 69364 Lyon Cedex 07, France.

γ_m/γ_g over a large pressure range means that the rare gas ‘prefers’ to enter the melt than remain in its parent fluid—because the fluid becomes so compressed. These findings are at variance with the common assumption that the Henry solubility is constant with increasing pressure and that the rare gas fluid can be considered ideal.

Some noble-gas solubility data exists for various silicate liquids at high pressure^{2,3}, but only recently was the 10–100-kbar range investigated^{4,5,13}. In Fig. 2 we present the Ar data obtained by ref. 5 with Fe-free synthetic haplogranitic and tholeiitic melts. These data are remarkably well described by our theoretical curves, which further predict a concentration maximum. The levelling off was interpreted⁵ as a saturation effect whereby the host sites for Ar atoms in melt all become occupied. In contrast, for liquid state theory, there are no predefined holes in a liquid (or a dense fluid), but the density fluctuations at a given pressure govern the probability of accommodating a solute of given size. Thus, the higher the pressure, the larger the density, and the smaller the solubility parameter (Fig. 1). The quasi-linear increase of Ar concentration with pressure is mimicked by the Henry law (dotted lines in Fig. 2) because of the concurrent compaction of melt and gas. If the rare-gas fluid was assumed ideal (dashed-dotted curves), the behaviour would be completely different and the quasi-linear regime restricted to much lower pressures.

Our ultimate goal is modelling the noble-gas fractionation between vesicles and melt for mid-ocean ridge basalts (MORBs). During CO₂ exsolution from a magma rising under ridges, the noble gases redistribute between bubbles and melt¹⁴ according to the mass conservation law:

$$C_0^i = C_v^i + C_m^i \quad (4)$$

where C_0^i (cm³ STP g⁻¹ or p.p.m.) is the concentration of rare gas ‘i’ in magma before vesiculation, C_v^i is its concentration in the vesicles

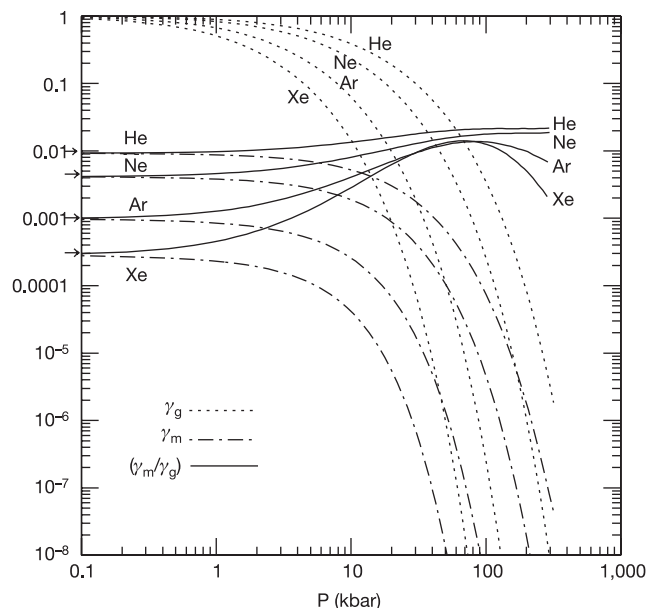


Figure 1 | Calculated solubility parameters of pure noble gases in a tholeiitic melt at 1,673 K as a function of pressure. The basalt composition is the same as that investigated by ref. 12; its density at one bar was deduced from its chemical composition using the partial molar volumes of oxide components⁹. The diameter of the hard spheres describing the basalt ($d = 3.18$ Å) was deduced by fitting high-pressure data from the literature²². The hard-sphere diameters of the noble gases ($d = 1.86, 2.21, 2.91$ and 3.515 Å for He, Ne, Ar and Xe respectively) were deduced from fits of their P - V - T data⁸ using the equation of state for hard spheres. The arrows indicate the values of the solubility parameter γ_m of the noble gases in melt at one atmosphere (equation (3)), which corresponds to the solubility constant S measured by ref. 12. Henry's law would correspond to horizontal lines.

containing the CO₂ phase in equilibrium with melt and C_m^i its concentration in the vesiculated melt. Introducing into equation (4) the equilibrium conditions, at given T and P , between the chemical potentials of the noble gas in melt and the CO₂ phase enables us to express C_m^i and C_v^i in terms of γ_m^i , the solubility parameter of the rare gas in melt, γ_v^i that in the CO₂ phase, and vesicularity $V^* = V_g/(V_g + V_m)$ (V_g , gas volume, V_m , melt volume). All these quantities depend on pressure and have to be calculated self-consistently (see Supplementary Information). The evolution of V^* with pressure in an ascending tholeiitic melt is shown in Fig. 3 for different initial conditions.

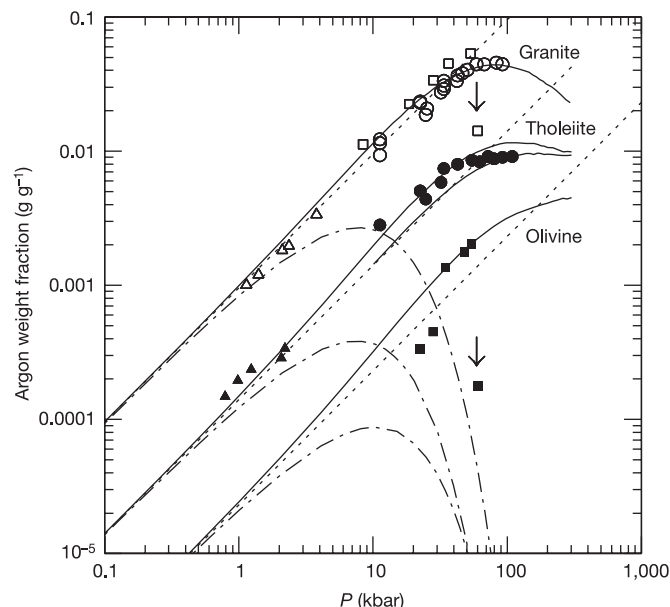


Figure 2 | Calculated argon concentration (weight fraction) for pure argon in contact with haplogranite, tholeiite and olivine melts, as a function of pressure. Temperature is 2,000 K for haplogranite, 1,800 and 2,300 K (the full curve interrupted below 10 kbar) for tholeiite because experiments cover this temperature range, and 2,300 K for olivine. Melt densities at 1 bar were deduced from data in ref. 9. The hard-sphere diameter of haplogranite ($d = 3.27$ Å) was obtained by fitting the isothermal pressure dependence of the partial molar volume for SiO₂ in rhyolite-type silicate liquids recently determined by ref. 23. The hard-sphere diameter for tholeiite is assumed to be identical to that used in Fig. 1 ($d = 3.18$ Å), while the one describing olivine ($d = 2.82$ Å) is deduced from a fit of static compression data on ultrabasic silicate liquids²⁴. An uncertainty of about $\pm 10\%$ on pressure during the fitting procedure (the order of magnitude of the experimental uncertainties) leads to an uncertainty around $\pm 1\%$ on the hard-sphere diameter. Further, a change in the hard-sphere diameter of $\pm 1\%$ yields a variation in Ar concentration of $\pm 10\%$ at the maximum of concentration and less than $\pm 1\%$ in the linear region. The empty and solid circles represent the data of ref. 5 for their Fe-free synthetic haplogranite and tholeiite, respectively. Also shown for comparison in the intermediate pressure range (1–10 kbar) are some Ar data for rhyolite³ (empty triangles) and natural basalts³ (full triangles), whose compositions are very close to those investigated by ref. 5. In the case of olivine, our calculated argon solubility constant at 1 bar ($S^{\text{Ar}} = 1.8 \times 10^{-5}$ cm³ STP g⁻¹ bar⁻¹, see the dotted line) is between that measured in enstatite melts^{25,26} ($S^{\text{Ar}} \approx 2.0 \times 10^{-5}$ cm³ STP g⁻¹ bar⁻¹) and that measured in komatiite²⁷ ($S^{\text{Ar}} \approx 1.0 \times 10^{-5}$ cm³ STP g⁻¹ bar⁻¹). Nevertheless, we cannot reproduce the entire data set of ref. 13 (solid squares) and especially the abrupt, order-of-magnitude drop of Ar concentration at around 50 kbar (vertical arrow). In another study⁴ data for Ar in silica (empty squares) also exhibit an abrupt decrease at similar pressure. These data contradict those of ref. 5 for a highly silicic melt (haplogranite), and our own calculations for silica (not shown), which present a broad maximum as for the haplogranite. An explanation could be the occurrence of an abrupt change in melt structure induced or not by the presence of rare-gas atoms, a scenario not described by our model.

Among tholeiitic glasses from oceanic ridges, popping rocks are those presenting the largest vesicularity, the largest volatile concentration, the lowest $^4\text{He}/^{40}\text{Ar}$ ratio and a regular vesicle size distribution^{15–17}. Therefore, they are often considered undegassed. Hence, we used, as initial CO_2 abundance, the concentration of 0.75 wt% estimated for the popping rock 2IID43 (ref. 18), corresponding to saturation at ~ 10 kbar (~ 35 km below sea level). Calculated vesicularity at eruption ($\sim 3,300$ m below sea level) is $\sim 17\%$ (Fig. 3), in good agreement with measured values¹⁵ (vesicle size distribution is not considered). Assuming an initial $^4\text{He}/^{40}\text{Ar}$ ratio of 1.5 (refs 15, 19), our model predicts at eruption a ratio of 13.5 in melt and 1.45 in vesicles, with 96% of the He and 99.55% of the Ar atoms in vesicles (the whole sample being dominated by vesicles). These results confirm that popping rocks represent a single equilibrium vesiculation with no vesicle loss.

In general, MORBs present a larger variation of their $^4\text{He}/^{40}\text{Ar}$ ratio (~ 1 –100), a ratio negatively correlated with Ar concentration¹ and vesicularity. As suggested in ref. 1, elemental fractionation between melt and vesicles could be responsible for these trends if vesiculation is followed by vesicle loss^{1,15}, possibly in several stages, provided some pressure effect acts on noble-gas solubility. We have thus evaluated the $^4\text{He}/^{40}\text{Ar}$ ratio in melt, vesicles and the bulk magma, when, during ascent, successive vesiculation stages start at various depths and finish with complete vesicle loss (except the last one). Again, the initial He, Ar and CO_2 concentrations are from 2IID43, and the first vesiculation starts at a depth of ~ 35 km ($P_{\text{sat}} = 10$ kbar).

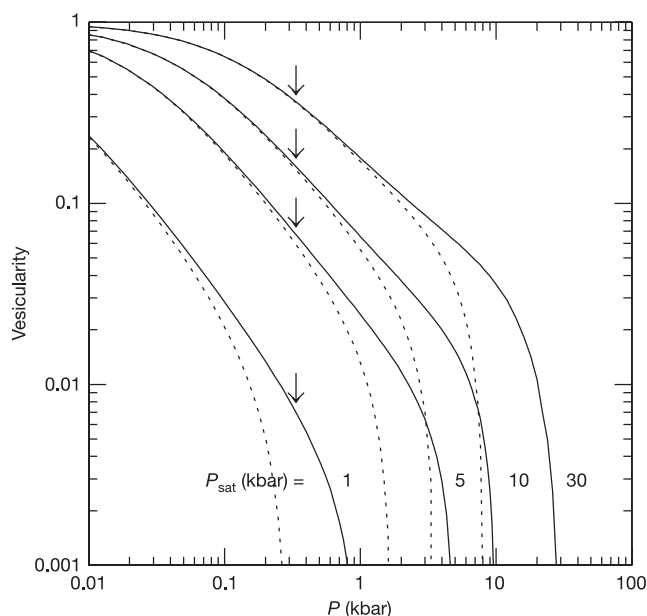


Figure 3 | Evolution of the vesicularity of a MORB with pressure during magma ascent. Assuming that the vesiculation stage occurs when the pressure in magma equals the saturation pressure of CO_2 , four initial conditions are presented, namely $P_{\text{sat}} = 1, 5, 10$ and 30 kbar (full curves). The dotted curves correspond to exsolution beginning only after a certain degree of super-saturation is reached (here four times the CO_2 saturation value at P_{sat}). The vertical arrows indicate the value of vesicularity at eruption, $3,300$ m under sea level. For $P_{\text{sat}} = 10$ kbar, this vesicularity is around 17% and is virtually independent of supersaturation (a fact corroborated by a recent experimental study, in which vesicularity at a given pressure is almost independent of decompression rate²⁸). Vesicularity evaluation needs to know the evolution of the CO_2 concentration with P . Our model interpolates very well the existing solubility data on CO_2 in tholeiitic melts over the entire pressure range investigated up to now²⁹ (see Supplementary Information). Note that higher vesicularity is predicted when eruption depth is shallower than $3,300$ m.

In Fig. 4 we present the calculated values at eruption ($\sim 3,300$ m) of the $^4\text{He}/^{40}\text{Ar}$ ratio versus Ar concentration and vesicularity for three successive stages. The second vesiculation starts at various depths between 35 km and eruption. The third vesiculation starts between the depth of the second stage and eruption, and is shown for a second stage at 4 kbar (~ 16 km) and 2 kbar (~ 8 km). Fractionation differs significantly in melt, vesicles and the bulk system. The bulk $^4\text{He}/^{40}\text{Ar}$ ratio becomes higher than ten only when the third vesiculation is relatively shallow ($P \leq 4$ kbar). The agreement with recent data on MORB glasses, mostly obtained by heating (ref. 1), is excellent, supporting the conclusion that several stages of vesiculation and vesicle loss are responsible for the $^4\text{He}/^{40}\text{Ar}$ variations in MORBs.

Our calculations explain why the data points do not follow a single master curve but are distributed over a large area. The key is that the final values depend on the depths of the successive vesiculation stages, a feature missed when assuming pressure-independent, Henry-type solubilities. The noble-gas elemental fractionation in MORBs keeps a record of these events. To access this record, we need accurate analyses of both melt and vesicles, and high-pressure experiments to obtain noble-gas solubility as a function of pressure.

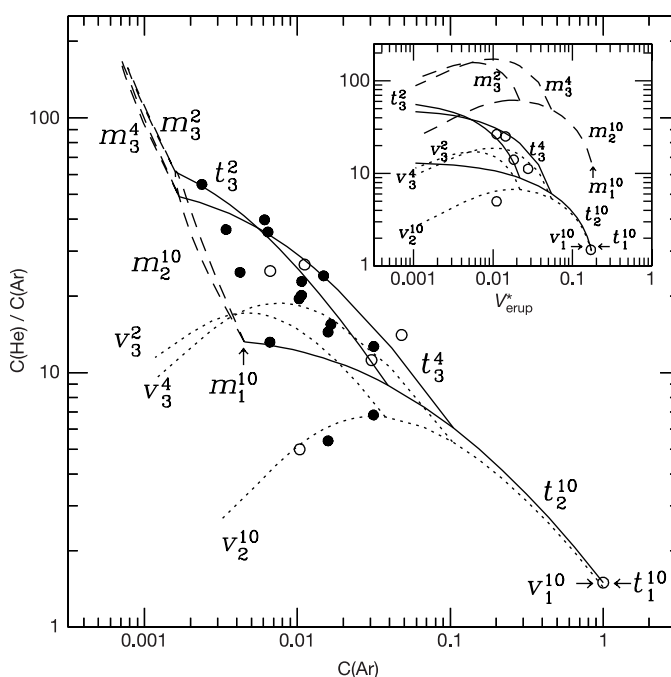


Figure 4 | Calculated $^4\text{He}/^{40}\text{Ar}$ ratio compared with measurements for MORBs. Shown are the variations of the $^4\text{He}/^{40}\text{Ar}$ ratio at eruption, $3,300$ m under sea level, with the Ar concentration (or vesicularity at eruption, V_{erup} in the inset) in a representative MORB after 1, 2 and 3 vesiculation and vesicle-loss stages; third-stage vesicles are not lost and the major gas is CO_2 . Concentrations are normalized to the values found in the 2IID43 popping rock¹⁹. The first vesiculation occurs at $P_{\text{sat}} = 10$ kbar at normal saturation; the influence of a possible supersaturation is marginal and not shown. The full curves (indexed by t_i^j) correspond to the bulk magma (melt + vesicles), the dotted curves (indexed by v_i^j) to vesicles and the dashed curves (indexed by m_i^j) to melt. Subscript i indicates the number of vesiculation stages and superscript j gives the pressure (in kbar) at which vesiculation stage ($i - 1$) occurred. For example, v_3^4 means that the corresponding curve gives the running values (with P) of the Ar concentration and $^4\text{He}/^{40}\text{Ar}$ ratio in vesicles after three vesiculations: the first stage occurs at 10 kbar, the second one at 4 kbar and the third one at various depths between 4 and 0.33 kbar. Dots are data from the literature mostly obtained by heating as used in ref. 1. Empty circles show those MORBs for which vesicularity was measured. Most data can be explained with three stages, the third one occurring between 4 and 1 kbar.

Degassing at ridges starts as deep as 35 km in the mantle where the first vesicles, filled with dense, supercritical CO₂, appear in ascending melts. Vesicles grow both by CO₂ feeding and expansion. Generally, magma slows down or stops at constrictions in conduits and, eventually, in a magma chamber; there, vesicles are lost. Most MORBs are thus quite strongly degassed. Only in some rare cases, such as popping rocks, does magma not stop during ascent, so such samples are representative of undegassed magma. Among consequences, the CO₂ outgassing flux is probably high¹⁵, and the upper-mantle ³He concentration may be higher than many estimates, possibly having a strong influence on our vision of the interior of the Earth^{20,21}.

Received 6 July 2004; accepted 7 April 2005.

- Sarda, P. & Moreira, M. Vesiculation and vesicle loss in mid-ocean ridge basalt glasses: He, Ne, Ar elemental fractionation and pressure influence. *Geochim. Cosmochim. Acta* **66**, 1449–1458 (2002).
- White, B. S., Brearley, M. & Montana, A. Solubility of argon in silicate liquid at high pressures. *Am. Mineral.* **74**, 513–529 (1989).
- Carroll, M. R. & Stolper, E. M. Noble gas solubilities in silicate melt and glasses: new experimental results for argon and the relationship between solubility and ionic porosity. *Geochim. Cosmochim. Acta* **57**, 5039–5051 (1993).
- Chamorro-Pérez, E., Gillet, Ph. & Jambon, A. Argon solubility in silicate melts at very high pressures. Experimental set-up and preliminary results for silica and anorthite melts. *Earth Planet. Sci. Lett.* **145**, 97–107 (1996).
- Schmidt, B. C. & Keppler, H. Experimental evidence for high noble gas solubilities in silicate melts under mantle pressures. *Earth Planet. Sci. Lett.* **195**, 277–290 (2002).
- Andersen, H. C., Weeks, J. D. & Chandler, D. Relationship between the hard-sphere fluid and fluids with realistic repulsive forces. *Phys. Rev. A* **4**, 1597–1605 (1971).
- Carnahan, N. F. & Starling, K. E. Equation of state for nonattracting rigid spheres. *J. Chem. Phys.* **51**, 635–636 (1969).
- Song, Y. & Mason, E. A. Statistical-mechanical theory of a new analytical equation of state. *J. Chem. Phys.* **91**, 7840–7853 (1989).
- Bottinga, Y., Richet, P. & Weill, D. F. Calculation of the density and thermal expansion coefficient of silicate liquids. *Bull. Minéral.* **106**, 129–138 (1983).
- Reiss, H., Frisch, H. L., Helfand, E. & Lebowitz, J. L. Aspects of the statistical thermodynamics of real fluids. *J. Chem. Phys.* **32**, 119–124 (1960).
- Guillot, B. & Guissani, Y. The solubility of rare gases in fused silica: a numerical evaluation. *J. Chem. Phys.* **105**, 255–270 (1996).
- Jambon, A., Weber, H. & Braun, O. Solubility of He, Ne, Ar, Kr, Xe in a basalt melt in the range 1250–1600 °C. Geochemical implications. *Geochim. Cosmochim. Acta* **50**, 401–408 (1986).
- Chamorro-Pérez, E., Gillet, Ph., Jambon, A., Badro, J. & McMillan, P. Low argon solubility in silicate melts at high pressure. *Nature* **393**, 352–355 (1998).
- Bottinga, Y. & Javoy, M. MORB degassing: bubble growth and ascent. *Chem. Geol.* **81**, 255–270 (1990).
- Sarda, P. & Graham, D. Mid-ocean ridge popping rocks: implications for degassing at ridge crests. *Earth Planet. Sci. Lett.* **97**, 268–289 (1990).
- Javoy, M. & Pineau, F. The volatiles record of a “popping” rock from the mid-atlantic ridge at 14 °N: chemical and isotopic composition of gas trapped in the vesicles. *Earth Planet. Sci. Lett.* **107**, 598–611 (1991).
- Burnard, P., Graham, D. & Turner, G. Vesicle-specific noble gas analyses of ‘popping rock’: implications for primordial noble gases in the Earth. *Science* **276**, 568–571 (1997).
- Graham, D. & Sarda, P. Reply to comment by T.M. Gerlach on ‘Mid-ocean ridge popping rocks: implications for degassing at ridge crests’. *Earth Planet. Sci. Lett.* **105**, 568–573 (1991).
- Moreira, M., Kunz, J. & Allègre, C. Rare gas systematics in popping rock: isotopic and elemental compositions in the upper mantle. *Science* **279**, 1178–1181 (1998).
- Allègre, C. J., Staudacher, T. & Sarda, P. Rare gas systematics: formation of the atmosphere, evolution and structure of the Earth’s mantle. *Earth Planet. Sci. Lett.* **81**, 127–150 (1987).
- Ballentine, C. J., van Keken, P. E., Porcelli, D. & Hauri, E. H. Numerical models, geochemistry and the zero-paradox noble gas mantle. *Phil. Trans. R. Soc. Lond. A* **360**, 2611–2631 (2002).
- Rigden, S. M., Ahrens, T. J. & Stolper, E. M. Densities of liquid silicates at high pressures. *Science* **226**, 1071–1074 (1984).
- Gaetani, G. A., Asimow, P. D. & Stolper, E. Determination of the partial molar volume of SiO₂ in silicate liquids at elevated pressures and temperatures: a new experimental approach. *Geochim. Cosmochim. Acta* **62**, 2499–2508 (1998).
- Agee, C. B. & Walker, D. Static compression and olivine flotation of ultrabasic silicate liquid. *J. Geophys. Res.* **93**, 3437–3449 (1988).
- Kirsten, T. Incorporation of rare gases in solidifying enstatite melts. *J. Geophys. Res.* **73**, 2807–2810 (1968).
- Shibata, T., Takahashi, E. & Matsuda, J.-I. Solubility of neon, argon, krypton and xenon in binary and ternary silicate systems: a new view on noble gas solubility. *Geochim. Cosmochim. Acta* **62**, 1241–1253 (1998).
- Carroll, M. R. & Draper, D. S. Noble gases as trace elements in magmatic processes. *Chem. Geol.* **117**, 37–56 (1994).
- Mourtada-Bonnefoi, C. C. & Laporte, D. Kinetics of bubble nucleation in a rhyolitic melt: an experimental study of the effect of ascent rate. *Earth Planet. Sci. Lett.* **218**, 521–537 (2004).
- Pan, V., Holloway, J. R. & Hervig, R. L. The pressure and temperature dependence of carbon dioxide solubility in tholeiitic basalt melts. *Geochim. Cosmochim. Acta* **55**, 1587–1595 (1991).

Supplementary Information is linked to the online version of the paper on www.nature.com/nature.

Acknowledgements We thank P. Burnard for sharing data and discussions, and P. Richet for discussions and encouragements.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.S. (sarda@geophy.geol.u-psud.fr).

Generation time and temporal scaling of bird population dynamics

Bernt-Erik Sæther¹, Russell Lande², Steinar Engen³, Henri Weimerskirch⁴, Magnar Lillegård³, Res Altwegg⁵, Peter H. Becker⁶, Thomas Bregnballe⁷, Jon E. Brommer⁸, Robin H. McCleery⁹, Juha Merilä⁸, Erik Nyholm¹⁰, Wallace Rendell¹¹, Raleigh R. Robertson¹², Piotr Tryjanowski¹³ & Marcel E. Visser¹⁴

Theoretical studies have shown that variation in density regulation strongly influences population dynamics¹, yet our understanding of factors influencing the strength of density dependence in natural populations still is limited². Consequently, few general hypotheses have been advanced to explain the large differences between species in the magnitude of population fluctuations^{3–6}. One reason for this is that the detection of density regulation in population time series is complicated by time lags induced by the life history of species^{7,8} that make it difficult to separate the relative contributions of intrinsic and extrinsic factors to the population dynamics. Here we use population time series for 23 bird species to estimate parameters of a stochastic density-dependent age-structured model. We show that both the strength of total density dependence in the life history and the magnitude of environmental stochasticity, including transient fluctuations in age structure, increase with generation time. These results indicate that the relationships between demographic and life-history traits in birds^{9,10} translate into distinct population dynamical patterns that are apparent only on a scale of generations.

Analyses of density dependence in natural populations are usually based on autoregression of population time series in which regression coefficients are assumed to represent both direct and delayed density dependence¹¹. This approach neglects the basic fact that the life history can produce time lags in the population dynamics⁸ that will wrongly be interpreted as delayed density dependence^{12,13}. In the Soay sheep (*Ovis aries*) lagged responses in life history were shown to be important for explaining temporal variation in population fluctuations¹⁴. In a simple deterministic model with no age structure, de Kroon *et al.*¹⁵ defined the strength of density dependence γ as the negative elasticity of the population growth rate λ with respect to changes in population size N , evaluated at the carrying capacity K , $\gamma = -(\partial \ln \lambda / \partial \ln N)_K$. Lande *et al.*¹⁶ extended this approach to show that in a generalized life-history model structured by age and dependent on density, total density dependence in the life history, D , should be defined as the negative elasticity of the population growth rate per generation, λ^T , with respect to the change in the size of the adult population when fluctuating around the carrying capacity, so that

$$D = - \left(T \frac{\partial \ln \lambda}{\partial \ln N} \right)_K \quad (1)$$

where T is the generation time. Thus, the rate of return to equilibrium then becomes $\gamma = D/T$. This definition facilitates a comparison of the strength of density dependence between species with different life-history characteristics.

Here we use the general definition of density dependence in equation (1) to model the stochastic density-dependent population dynamics of different bird species. For simplicity we assume that the expected adult annual survival and fecundity rates are independent of age. Furthermore, it is assumed that density dependence is exerted by the adult fraction of the population on any combination of juvenile and adult vital rates, which encompasses general features of the dynamics of many bird populations¹⁷. Finally, deviations of the adult population at time t from equilibrium, $x(t) = N(t) - K$, are expected to be small or moderate. In our data set, the coefficient of variation in the time series is generally less than 30%, which has been shown^{5,18} to fit our theoretical approach well. On the basis of these assumptions, we obtain¹⁶ a linearized autoregressive model with time delays from 1 to α years, namely

$$x(t) = \sum_{i=1}^{\alpha} b_i x(t-i) + \omega(t) \quad (2)$$

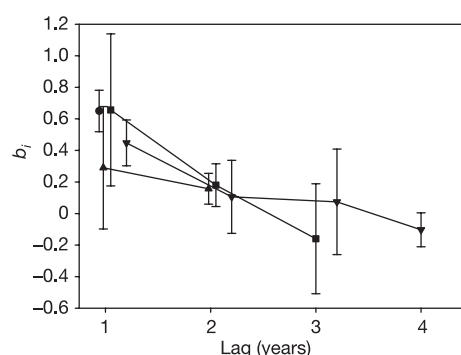


Figure 1 | Autoregression coefficients b_i for different lags in the population dynamics in relation to variation in age at maturity. Circles represent species that mature at 1 year, triangles age at maturity at 2 year, squares age at maturity at 3 years and reversed triangles species that mature at 4 years or older. Results are means $\pm$ s.d.

¹Department of Biology, Norwegian University of Science and Technology, N-7491 Trondheim, Norway. ²Department of Biology 0116, University of California San Diego, La Jolla, California 92093, USA. ³Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway. ⁴CEBC - CNRS, FR-79360 Villiers en Bois, France. ⁵Department of Biology, University of Victoria, PO Box 3020, Victoria, British Columbia V8W 3N5, Canada. ⁶Institut für Vogelforschung, Vogelwarte Helgoland, An der Vogelwarte 21, D-26386 Wilhelmshaven, Germany. ⁷National Environmental Research Institute, Department of Coastal Zone Ecology, Kalø, Grenåvej 12, DK-8410 Rønde, Denmark. ⁸Department of Biological and Environmental Sciences, University of Helsinki, PO Box 65, FIN-00014 Helsinki, Finland. ⁹Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. ¹⁰Umeå University, Department of Ecology and Environmental Sciences, S-90187 Umeå, Sweden. ¹¹Applied Sciences and Computing, Loyalist College, Wallbridge-Loyalist Road, PO Box 4200, Belleville, Ontario K8N 5B9, Canada. ¹²Department of Biology, Queen's University, Kingston, Ontario K7L 3N6, Canada. ¹³Department of Behavioural Ecology, Adam Mickiewicz University, Umultowska 89, PL-61614 Poznań, Poland. ¹⁴Netherlands Institute of Ecology, PO Box 40, NL-6666 ZG Heteren, The Netherlands.

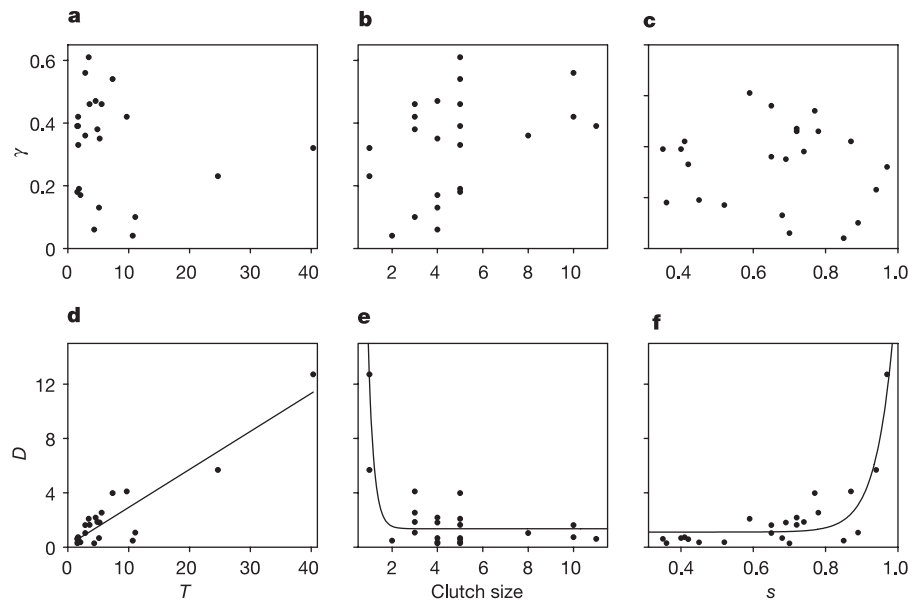


Figure 2 | Strength of density dependence in relation to life history variation in birds. The strength of density dependence in annual population

fluctuations, γ (a–c), and the total life history, D (d–f), in relation to generation time T (a, d), clutch size (b, e) and adult survival rate s (c, f).

where $\omega(t)$ is a noise term with a mean of zero and variance σ_ω^2 , describing environmental stochasticity, including transient fluctuations in age structure and autocorrelations due to long-term fluctuations in the biotic or abiotic environment.

In our data set the mean of the autoregression coefficients b_i for species with age at maturity $\alpha \geq 2$ decreased with time lag (Fig. 1), indicating that the effects of the previous years' population sizes on current population size decreased with time. However, these autoregression coefficients b_i do not directly reveal the strength of delayed density dependence because they depend on both life-history parameters and density dependence in the vital rates^{5,18}. For instance, in a species with $\alpha > 1$, with no density dependence in subadult or adult survival rates, b_1 equals the adult annual survival rate s . Similarly, in a species that matures at 1 year, if $b_1 = 0$ the population autocorrelations for all time lags will be zero, corresponding to a white-noise process for the population size, $N(t)$, that indicates strong density dependence. However, on the basis of the general definition of density dependence (equation (1)), there is a relationship¹⁶ between total density dependence in the life history and the autoregression coefficients:

$$(1-s)D = 1 - \sum_{i=1}^{\alpha} b_i \quad (3)$$

Using this theoretical framework, we find a large interspecific variation in the density dependence in bird population dynamics. The strength of density dependence varied from $\gamma = 0.04$ in the South Polar skua (*Catharacta macrormicki*) to $\gamma = 0.61$ in the Eurasian sparrowhawk (*Accipiter nisus*), resulting in differences in return times to equilibrium from about 1.7 to 22.5 years. These density-dependent effects were independent of generation time (Fig. 2a; correlation coefficient = -0.15 , $P > 0.5$) as well as clutch size (Fig. 2b; correlation coefficient = 0.40 , $P = 0.06$) and adult survival rate (Fig. 2c; correlation coefficient = -0.09 , $P > 0.69$). As a consequence, the stationary variance in the time series σ_N^2 was also independent of life history (the absolute value of the correlation coefficients between σ_N^2 and T , clutch size, s and α were all less than 0.10; $P > 0.7$).

Our analyses reveal that the magnitude of annual fluctuations in size of age-structured bird populations cannot be predicted from estimates of generation time or from a knowledge of life-history traits such as clutch size and adult survival rate. However, previous

comparative analyses have shown that many avian demographic traits such as clutch size and age at maturity scale closely with adult life span^{9,19,20}. As expected from this, we find that several features of population dynamics measured on a time scale of generations can be predicted from life-history characteristics. The strength of total density dependence in the life history, D , increased with generation time T (Fig. 2d; correlation coefficient = 0.92,

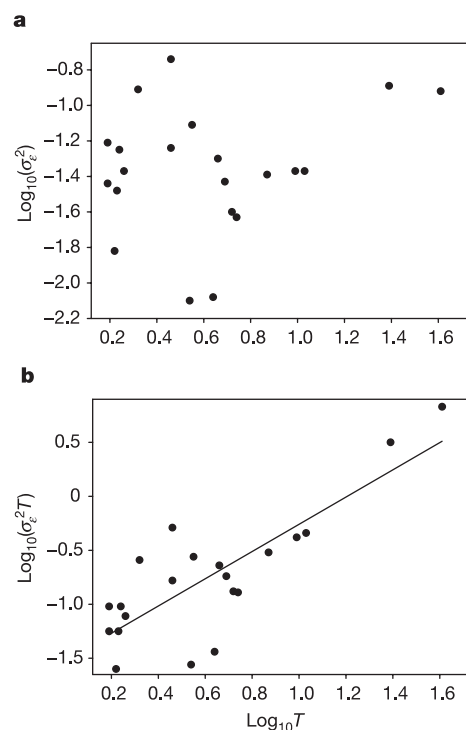


Figure 3 | Residual variation in avian population fluctuations after accounting for density dependence in relation to generation time T . a, The residual variance in a first-order process, σ_ϵ^2 , describing environmental stochasticity and transient fluctuations in age structure as well as long-term autocorrelations in the environment. b, The total residual variance $\sigma_\epsilon^2 T$ over a period of one generation.

$n = 23$, $P < 0.001$) and adult survival rate (Fig. 2f; correlation coefficient = 0.62, $n = 23$, $P = 0.002$) but decreased with clutch size (Fig. 2e; correlation coefficient = -0.45 , $n = 23$, $P = 0.03$). Although this relationship was strongly influenced by the two long-lived seabird species southern fulmar (*Fulmarus glacialisoides*) and lesser snow-petrel (*Pagodroma nivea*), a significant relationship was still present after excluding those two species from the analyses involving T (correlation coefficient = 0.47, $n = 21$, $P = 0.034$) and adult survival rate (correlation coefficient = 0.57, $n = 21$, $P = 0.007$), whereas no significant effect was present for clutch size (correlation coefficient = -0.20 , $n = 21$, $P > 0.1$). An increase in strength of density dependence with longevity when the censuses were taken at intervals of one generation has also been previously recorded in British birds²¹. This implies that the effect on the population growth rate per generation of a change in population size was larger for long-lived species than for short-lived species. Consequently, the rate of return to equilibrium measured in generations decreases with generation time T (correlation coefficient of $\log_{10}$ -transformed values = -0.73 , $P < 0.001$, $n = 23$).

To compare the residual variation in the population process we must account for interspecific variation in age at maturity that will cause differences in the lag-structure of the population dynamics¹⁶. We first estimate (see Methods) the variance in the stationary distribution of the population sizes σ_N^2 in our model (equation (2)) and then calculate the variance in the noise of a first-order process with a single time lag of one year, σ_e^2 , that will give the same stationary variance in population size as in the full model. The variance of this white noise process for species with age at maturity larger than 1 year should be approximately equal to the environmental variance (R.L., S.E., B.-E.S., and T. N. Coulson, unpublished observations). In our data set $\log_{10}\sigma_e^2$ was independent of $\log_{10} T$ (Fig. 3a, correlation coefficient = 0.24, $P > 0.3$, $n = 23$). In contrast, there was a highly significant linear increase in $\log_{10}(\sigma_e^2 T)$ (Fig. 3b; correlation coefficient = 0.82, $P < 0.001$, $n = 21$; correlation coefficient = 0.56, $n = 19$, $P = 0.013$ after omitting the southern fulmar and lesser snow-petrel). Furthermore, the environmental variance for this process per generation ($\sigma_e^2 T$) was closely related to life-history characters; that is, $\sigma_e^2 T$ decreased with clutch size (correlation coefficient of $\log_{10}$ -transformed values = -0.71 , $P < 0.001$, $n = 21$; no longer significant after removing the two extreme values, correlation coefficient = -0.25 , $n = 19$, $P > 0.3$) but increased with adult survival rate (correlation coefficient of $\log_{10}$ -transformed values = 0.70, $P < 0.001$, $n = 21$; correlation coefficient = 0.63, $n = 19$, $P = 0.004$ after omitting the two outliers). This indicates that environmental stochasticity per generation might be greater for long-lived species than for species with short life expectancies. In contrast, interspecific differences in avian demographic stochasticity per generation are independent of life-history variation²².

Comparative studies have shown that density dependence strongly influences population dynamics, and especially that of small passerine birds¹⁷. Increasing evidence suggests that density dependence also occurs frequently in long-lived species. For instance, in long-lived seabirds competition for breeding space or food can cause density dependence in their population dynamics²³. Similarly, environmental fluctuations strongly influence variation in population size²⁴ and age structure^{25,26} of many long-lived species and often generate lagged effects in the population dynamics²⁷. Although the effects of density dependence and environmental fluctuations may be small in long-lived species on a yearly basis¹⁹, their accumulated influence on the population growth rate over one generation may still be large. Despite large differences between species in sensitivity of population growth rate per generation to changes in population density and environmental effects (Figs 2d and 3b), the lack of trend in Fig. 3a indicates that life histories might have evolved to produce comparable environmental variance in both short-lived and long-lived

species, with similar impacts on their rates of population growth in the long run⁵.

Our results indicate that understanding population dynamical patterns in birds might require accounting for generation time. Because many bird species are long-lived it will be necessary for the progress of basic ecology to secure continuity of long-term demographical studies. For population viability analyses of rare or endangered species, in which it is often not feasible to obtain detailed demographic data, the development of empirical scaling relationships between stochastic population parameters such as demographic and environmental variances and generation time may provide useful information for extinction risk assessment as well as conservation and restoration planning.

METHODS

Data. This study is based on 23 time series longer than 15 years. Estimates of population parameters based on time series analyses are strongly influenced by the precision in the population estimates²⁸. We therefore included only studies in which uncertainties in population estimates are negligible compared with the environmental variance; that is, studies based on total counts of nests or colour-ringed birds²⁹. The average population size $\bar{N}$ should be much larger than σ_d^2/σ_e^2 to avoid the influence of demographic stochasticity⁵, where σ_d^2 is the environmental variance. No estimates of the demographic variance σ_d^2 were available for many of the species. Because demographic stochasticity has the largest influence on the population dynamics of small passerines with estimated values of σ_d^2 about 0.50 (ref. 22), we included only time series of passerines with mean values of more than 50 individuals.

Demographic data were extracted from studies of individual species (see Supplementary Information). Age at maturity α refers to the age at which regular breeding of females first occurred. Generation time T was calculated⁵ as $T = \alpha + [s/(1-s)]$, where s is the expected adult survival rate. Survival estimates were based either on capture–recapture analysis or calculated as the return rate of individually known adults from one breeding season to another.

Estimation of parameters. As described in ref. 16, the maximum-likelihood estimates of the autoregression coefficients b_i in equation (2) are found by inverting the Yule–Walker equations³⁰ for the time series. These estimators are biased because population size at a given time enters the regression as both a dependent variable and an independent variable¹³. This bias can be estimated and corrected by using computer simulations¹⁶. Those simulations also can be used for significance testing and calculation of confidence intervals of the estimates¹⁶.

Calculation of variance in the noise. For a stationary time series such as equation (2) the autocovariances $C(\tau) = \text{cov}(N_t, N_{t+\tau})$, $\tau = 0, 1, \dots$, are determined by the Yule–Walker equations³⁰

$$C(j) - \sum_{i=1}^{\alpha} b_i C(j-i) = \begin{cases} \sigma_e^2 & \text{for } j=0 \\ 0 & \text{for } j>0 \end{cases}$$

where the variance in the noise σ_e^2 is estimated by maximum likelihood³⁰ and $C(0) = \sigma_N^2$. Using the fact that $C(\tau) = C(-\tau)$, this is a set of linear equations that can be solved with respect to the variance in the stationary distribution $C(0)$ as well as $C(1)$, $C(2)$, We calculate the residual variance of a first-order process with the same stationary variance as in equation (2), namely $\sigma_e^2 = C(0)(1 - b_1^2)$, provided that for this first-order process $|b_1| < 1$.

Received 11 March; accepted 20 April 2005.

- May, R. M. Simple mathematical models with very complicated dynamics. *Nature* **261**, 459–467 (1976).
- Sibly, R. & Hone, J. in *Wildlife Population Growth Rates* (eds Sibly, R. M., Hone, J. & Clutton-Brock, T. H.) 11–40 (Cambridge Univ. Press, Cambridge, 2003).
- Pimm, S. L. *The Balance of Nature?* (Univ. of Chicago Press, Chicago, 1991).
- Ariño, A. & Pimm, S. L. On the nature of population extremes. *Evol. Ecol.* **9**, 429–443 (1995).
- Lande, R., Engen, S. & Sæther, B.-E. *Stochastic Population Dynamics in Ecology and Conservation* (Oxford Univ. Press, Oxford, 2003).
- Hanski, I. Density dependence, regulation and variability in natural populations. *Phil. Trans. R. Soc. Lond. B* **330**, 141–150 (1990).
- May, R. M. *Stability and Complexity in Model Ecosystems* (Princeton Univ. Press, Princeton, New Jersey, 1973).
- Nisbet, R. M. in *Structured Population Models in Marine, Terrestrial, and Freshwater Systems* (eds Tuljapurkar, S. & Caswell, H.) 89–118 (Chapman & Hall, New York, 1997).
- Charnov, E. L. A note on dimensionless life-histories for birds versus mammals. *Evol. Ecol.* **9**, 288–291 (1995).

10. Charnov, E. L. Evolution of life-history variation among species of altricial birds. *Evol. Ecol. Res.* **2**, 365–373 (2000).
11. Royama, T. *Analytical Population Dynamics* (Chapman & Hall, London, 1992).
12. Zeng, Z., Nowierski, R. M., Taper, M. L., Dennis, B. & Kemp, W. P. Complex population dynamics in the real world: modeling the influence of time-varying parameters and time lags. *Ecology* **79**, 2193–2209 (1998).
13. Bulmer, M. G. The statistical analysis of density dependence. *Biometrics* **31**, 901–911 (1975).
14. Coulson, T. et al. Age, sex, density, winter weather, and population crashes in Soay sheep. *Science* **292**, 1528–1531 (2001).
15. de Kroon, H., Plaisier, A., van Groenendaal, J. & Caswell, H. Elasticity: the relative contribution of demographic parameters to population growth rate. *Ecology* **67**, 1427–1431 (1986).
16. Lande, R. et al. Estimating density dependence from population time series using demographic theory and life-history data. *Am. Nat.* **159**, 321–332 (2002).
17. Newton, I. *Population Limitation in Birds* (Academic, San Diego, 1998).
18. Lande, R. in *Population viability analysis* (eds Beissinger, S. & McCullough, D. R.) 18–40 (Univ. of Chicago Press, Chicago, 2002).
19. Sæther, B.-E. & Bakke, Ø. Avian life history variation and contribution of demographic traits to the population growth rate. *Ecology* **81**, 642–653 (2000).
20. Bennett, P. M. & Owens, I. P. F. *Evolutionary Ecology of Birds* (Oxford Univ. Press, Oxford, 2002).
21. Holyoak, M. & Baillie, S. R. Factors influencing detection of density dependence in British birds. II. Longevity and population variability. *Oecologia* **108**, 54–63 (1996).
22. Sæther, B.-E. et al. Life-history variation predicts the effects of demographic stochasticity on avian population dynamics. *Am. Nat.* **164**, 793–802 (2004).
23. Croxall, J. P. & Rothery, P. in *Bird Population Studies. Relevance to Conservation and Management* (eds Perrins, C. M., Lebreton, J.-D. & Hiron, G. J. M.) 272–296 (Oxford Univ. Press, Oxford, 1991).
24. Sæther, B.-E., Sutherland, W. J. & Engen, S. Climate influences on avian population dynamics. *Adv. Ecol. Res.* **35**, 185–209 (2004).
25. Barbraud, C. & Weimerskirch, H. Climate and density shape population dynamics of a marine top predator. *Proc. R. Soc. Lond. B* **270**, 2111–2116 (2003).
26. Weimerskirch, H., Inchausti, P., Guinet, C. & Barbraud, C. Trends in bird and seal populations as indicators of a system shift in the Southern Ocean. *Antarct. Sci.* **15**, 249–256 (2003).
27. Thompson, P. M. & Ollason, J. C. Lagged effects of ocean climate on fulmar population dynamics. *Nature* **413**, 417–420 (2001).
28. De Valpine, P. & Hastings, A. Fitting population models incorporating process noise and observation error. *Ecol. Monogr.* **72**, 57–76 (2002).
29. Micol, T. & Jouventin, P. Long term population trends in seven Antarctic seabirds at Pointe Géologie (Terre Adélie). *Polar Biol.* **24**, 175–185 (2001).
30. Shumway, R. H. *Applied Statistical Time Series Analysis* (Prentice-Hall, Englewood Cliffs, New Jersey, 1988).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. P. Møller for help in organizing this database and for comments on the manuscript. This study was supported by the Research Council of Norway (Strategic University Programme in Conservation Biology), the Natural Sciences and Engineering Research Council of Canada, the Institut Polaire français-program 109 and the US National Science Foundation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.-E.S. (bernt-erik.sather@bio.ntnu.no).

Nicotine reinforcement and cognition restored by targeted expression of nicotinic receptors

U. Maskos¹, B. E. Molles¹, S. Pons¹, M. Besson^{1,2}, B. P. Guiard², J.-P. Guilloux², A. Evrard¹, P. Cazala³, A. Cormier¹, M. Mameli-Engvall¹, N. Dufour^{4†}, I. Cloëz-Tayarani¹, A.-P. Bemelmans^{4†}, J. Mallet⁴, A. M. Gardier², V. David³, P. Faure¹, S. Granon¹ & J.-P. Changeux¹

Worldwide, 100 million people are expected to die this century from the consequences of nicotine addiction¹, but nicotine is also known to enhance cognitive performance². Identifying the molecular mechanisms involved in nicotine reinforcement and cognition is a priority and requires the development of new *in vivo* experimental paradigms. The ventral tegmental area (VTA) of the midbrain is thought to mediate the reinforcement properties of many drugs of abuse. Here we specifically re-expressed the $\beta 2$ -subunit of the nicotinic acetylcholine receptor (nAChR) by stereotactically injecting a lentiviral vector into the VTA of mice carrying $\beta 2$ -subunit deletions^{3,4}. We demonstrate the efficient re-expression of electrophysiologically responsive, ligand-binding nicotinic acetylcholine receptors in dopamine-containing neurons of the VTA, together with the recovery of nicotine-elicited dopamine release and nicotine self-administration. We also quantified exploratory behaviours of the mice, and showed that $\beta 2$ -subunit re-expression restored slow exploratory behaviour (a measure of cognitive function) to wild-type levels, but did not affect fast navigation behaviour. We thus demonstrate the sufficient role of the VTA in both nicotine reinforcement and endogenous cholinergic regulation of cognitive functions.

Nicotinic acetylcholine receptors (nAChRs) are pentameric, ligand-gated ion channels abundant in the central nervous system⁵. Twelve neuronal subunits have been identified, designated $\alpha 2$ to $\alpha 10$, and $\beta 2$ to $\beta 4$, which potentially assemble in multiple combinations with a broad range of pharmacological and electrophysiological properties⁶. These receptors bind endogenous acetylcholine in the brain, but are also the target of nicotinic agonists and antagonists.

The midbrain VTA is considered the principal brain region mediating the reinforcement properties of multiple drugs of abuse, including nicotine, but its precise contribution is still challenged⁷. To further understand the specific role of $\beta 2$ -containing nAChRs (henceforth $\beta 2^*$ -nAChRs) in mediating the effects of nicotine and endogenous acetylcholine on reinforcement and cognition⁸, we selectively re-expressed the $\beta 2$ -subunit in the VTA of $\beta 2$ -knockout ($\beta 2^{-/-}$) mice.

Our approach was to generate a lentiviral expression vector (see Methods and Supplementary Information) containing a bi-cistronic cassette⁹ simultaneously expressing the $\beta 2$ -subunit and enhanced green fluorescent protein (eGFP), to facilitate efficient detection of transduced cells. Lentivectors¹⁰ have the advantage of allowing high-titre production (10^9 transforming units per ml), genomic integration of the recombinant provirus, the use of specific promoters, and the absence of viral-associated proteins that could provoke an

immune response. Despite their growing use in gene therapy and animal models, no functional neurotransmitter receptor has been expressed *in vivo* using this system.

The efficacy of our system was demonstrated by confocal analysis of transduced mouse VTA and its dopamine- and γ -aminobutyric acid (GABA)-containing neurons (Fig. 1a). Receptor autoradiography on brain slices¹¹ revealed the presence of the re-expressed high-affinity nAChR at the injection site in the VTA (Fig. 1b and Supplementary Fig. S2), as well as in the mesolimbic projection and axon terminals in the nucleus accumbens. These projections originate from both dopamine- (85%) and GABA-containing (15%) projection neurons that have cell bodies situated in the VTA¹² (see Supplementary Fig. S3). The amount of signal in the VTA was comparable to wild-type mice, but the distribution of re-expressed binding sites within the VTA was not uniform, with higher binding levels near the injection locus (Fig. 1b). The recovery of [¹²⁵I]-epibatidine binding sites indicates that the re-expressed $\beta 2$ -subunits, which are synthesized in the cell bodies of the VTA, can partner with endogenous α -subunits to form binding sites, and are efficiently transported to terminal fields in the nucleus accumbens, as observed in the wild-type brain.

The critical issue then became the *in vivo* functionality of the re-expressed receptor. For the four *in vivo* experimental paradigms described below, three different experimental groups of male mice were generated: wild-type C57BL/6J mice injected with the eGFP-only lentivector (hereafter referred to as 'WT'), $\beta 2^{-/-}$ mice injected with the eGFP-only lentivector ('KO'), and $\beta 2^{-/-}$ mice with the $\beta 2$ + eGFP bi-cistronic vector ('VEC'). The $\beta 2^{-/-}$ mice used in the experiments had been backcrossed with wild-type C57BL/6J¹³ mice for 19 generations.

nAChR functionality was assessed by recording the effect of nicotine on the *in situ* electrophysiological activity of dopamine neurons in the VTA¹³. In WT mice, intravenous injection of $30 \mu\text{g kg}^{-1}$ of nicotine caused the well-established, rapid 1.5-fold increase in firing frequency that lasted nearly 10 min (ref. 14, Fig. 2a). KO mice produced either no response, or in a few cases, short duration responses, in agreement with previous *in vitro* observations on midbrain slices⁴. VEC mice showed responses to nicotine that were characterized by the same rapid increase in firing frequency observed in WT mice, confirming re-expression of functional nAChRs. However, unlike neurons from WT mice, this effect did not persist for more than 2 min on average. These data show that re-expression of the $\beta 2$ -subunit exclusively in the VTA (Fig. 1a) is sufficient to restore an effect of nicotine on dopamine neurons.

¹Unité Récepteurs et Cognition, CNRS URA 2182, Institut Pasteur, 25 rue du Dr Roux, 75724 Paris Cedex 15, France. ²Laboratoire de Neuropharmacologie EA3544, Faculté de Pharmacie, Université Paris-Sud, 92296 Châtenay-Malabry Cedex, France. ³Laboratoire de Neurosciences Cognitives, CNRS UMR 5106, Université de Bordeaux-I, Avenue des Facultés, 33405 Talence Cedex, France. ⁴Laboratoire de Génétique Moléculaire de la Neurotransmission et des Processus Neurodégénératifs (LGN), CNRS-UMR 7091, Hôpital de la Pitié-Salpêtrière, 83 Boulevard de l'Hôpital, 75013 Paris, France. [†]Present addresses: Isotopic Imaging, Biochemical and Pharmacological Unit, and Image Program, Service Hospitalier Frédéric Joliot, 91401 Orsay, France (N.D.); Laboratoire d'oculogénétique, Hôpital Ophtalmique Jules Gonin, 1004 Lausanne, Switzerland (A.-P.B.).

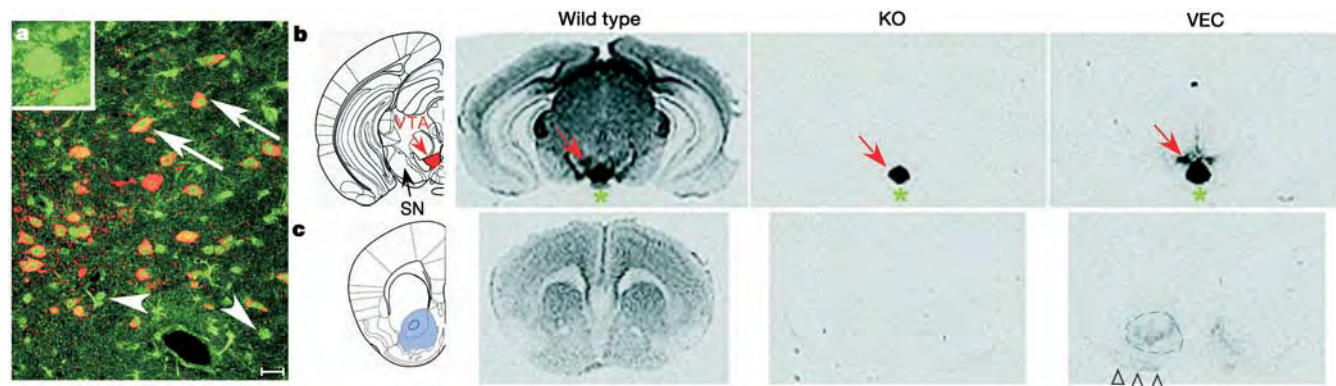


Figure 1 | Lentivector-mediated expression of eGFP and nAChR $\beta 2$ -subunit in the VTA. **a**, Confocal section of VTA neurons expressing eGFP (green). Tyrosine hydroxylase (TH, red) indicates dopamine neurons (arrows). GABA-containing neurons are TH-negative and are identified by their neuronal morphology (arrowheads and inset). **b**, [125 I]-epibatidine autoradiography, coronal sections at -3.4 mm from bregma. Arrows indicate the VTA (red) and the substantia nigra (SN, black). The

interpeduncular nucleus (which shows persistent epibatidine binding in $\beta 2^{-/-}$ mice¹¹) is directly above the green asterisk. **c**, [125 I]-epibatidine autoradiography in coronal sections at $+1.42$ mm from bregma. The nucleus accumbens is shown in blue (shading or outline); black arrowheads indicate re-expressed binding sites in the median forebrain bundle and olfactory tubercle.

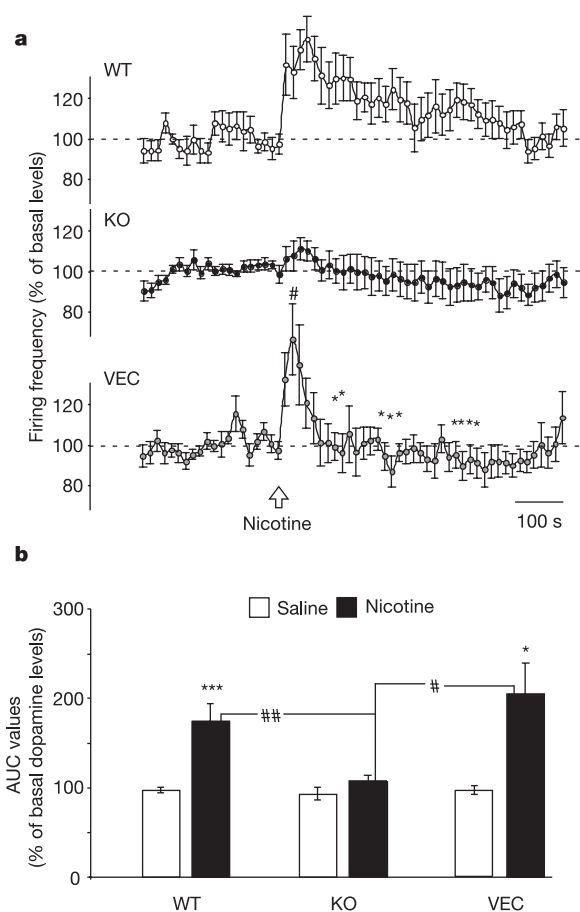


Figure 2 | Effects of nicotine *in vivo* on dopamine neuron firing and dopamine release. **a**, Responses in firing frequency of VTA neurons in WT ($n = 12$), KO ($n = 15$) and VEC ($n = 16$) mice, expressed as mean $\pm$ s.e.m. Asterisk, $P < 0.05$ between VEC and WT mice; hash symbol, $P < 0.05$ between VEC and KO mice. **b**, Extracellular dopamine levels in the nucleus accumbens shell, showing area under curve values (AUC values, mean $\pm$ s.e.m.) during a 2-h post-treatment period. AUC values are expressed as a percentage of dopamine levels of nicotine versus saline treatment. Nicotine administration increases dopamine release compared with saline in WT and VEC mice (Fisher test; single asterisk, $P < 0.05$; three asterisks, $P < 0.001$). Significant differences in release between nicotine-treated KO and WT and VEC mice are indicated (single hash symbol, $P < 0.01$; two hash symbols, $P < 0.001$).

Notably, the sustained (up to 10 min) increase in firing rate recorded in WT mice was not observed in the VEC mice. This suggests that the sustained firing of dopamine neurons in the VTA could be dependent on $\beta 2$ expression in other—excitatory—structures, such as glutamate-containing prefrontal cortex or ponto-tegmental afferents¹⁵. In addition, the precise nAChR subunit stoichiometry in the VEC mice might also be slightly different from WT mice, as most nAChR subunits are expressed in VTA neurons¹⁶.

To determine whether the re-expressed $\beta 2^*$ -nAChR receptor could mediate normal levels of nicotine-induced dopamine release, we used *in vivo* intracerebral microdialysis¹⁷ in awake, freely moving mice. In both VEC and WT mice, we observed a statistically significant increase in dopamine release in the nucleus accumbens following 1 mg kg^{-1} intraperitoneal nicotine injection (Fig. 2b). No nicotine-elicited dopamine release was observed in KO mice, as described for uninjected $\beta 2^{-/-}$ mice⁴. These data confirm a complete restoration of nicotine-elicited dopamine release from the nucleus accumbens by the re-expressed $\beta 2$ -subunit in the intact behaving animal.

To assess whether the nicotine-induced responses observed in the VTA and nucleus accumbens in VEC mice were sufficient to support nicotine reinforcement, we developed an intra-VTA nicotine self-administration paradigm in the mouse, as described for morphine self-administration¹⁸. Mice (7 WT, 10 KO, 9 VEC) were implanted and tested in a Y-maze. WT mice showed a clear nicotine-seeking behaviour from the second learning session, as measured both by increasing choice of the nicotine-reinforced arm (Fig. 3a) and decreasing self-injection latencies over time (Fig. 3b). In contrast, KO mice did not acquire nicotine self-administration behaviour. Their arm choice remained at chance level over the experiment, whereas latency for triggering injections gradually increased under nicotine sessions. This particular combination of parameters (no arm choice, long latency) is typically observed in non-reinforced animals¹⁸. As observed for WT mice, VEC mice also acquired intra-VTA nicotine self-administration behaviour and showed improved performance over learning sessions. However, they showed a delay in the acquisition of self-administration, differing from KO mice from the fourth learning session, and at the end of the experiment their discrimination performance was slightly lower than that observed for WT mice. As observed for WT mice, self-injection latency decreased over nicotine sessions in VEC mice, confirming nicotine-seeking behaviour (Fig. 3b). Therefore, we conclude that re-expression of $\beta 2^*$ -nAChR in the VTA is not only necessary⁴ but also sufficient for re-establishing sensitivity to nicotine reward in drug-naïve mice. The neuronal structures and molecules involved in nicotine addiction

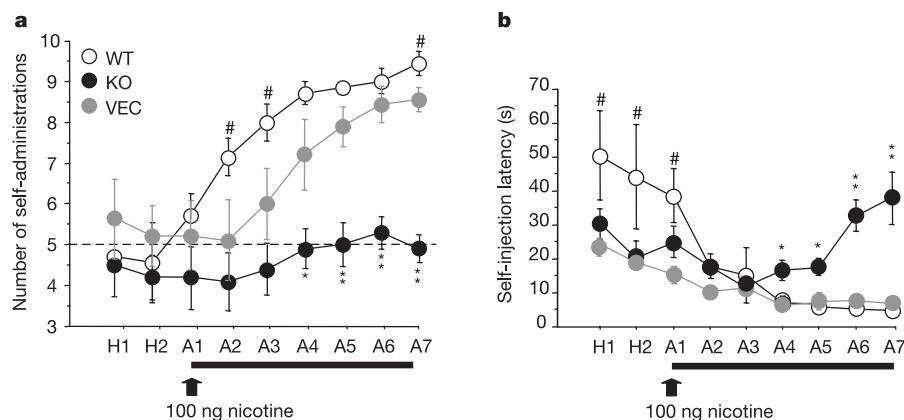


Figure 3 | Intra-VTA nicotine self-administration. **a**, Number of self-administrations per daily session of 10 trials, expressed as mean $\pm$ s.e.m. Days 1 and 2 consisted of habituation (H) trials, and days 3–9 involved nicotine self-administration trials (A1–7), with 100 ng nicotine (as salt) per self-administered dose. ANOVA revealed a significant group effect ($F_{2,22} = 8.315$, $P < 0.01$), session effect ($F_{8,176} = 17.907$, $P < 0.0001$) and group $\times$ session interaction ($F_{16,176} = 3.353$, $P < 0.0001$). Significant differences were identified by Fisher's test: single asterisk, $P < 0.05$ between KO and VEC mice; two asterisks, $P < 0.001$ between KO and VEC mice; hash symbol, $P < 0.05$ between WT and VEC mice. **b**, Injection latency

(mean $\pm$ s.e.m.) decreased drastically in WT and VEC mice, confirming expression of nicotine-seeking behaviour. Injection latency increased in KO mice. ANOVA revealed a significant group effect ($F_{2,22} = 4.637$, $P < 0.02$), a significant session effect ($F_{8,176} = 10.771$, $P < 0.0001$) and a strong group $\times$ session interaction ($F_{16,176} = 7.609$, $P < 0.0001$). Fisher post-hoc tests showed that mean self-injection latency between VEC and KO mice is significantly different from the fourth session (A4) onwards (single asterisk, $P < 0.05$; two asterisks, $P < 0.001$; hash symbol denotes a significant difference ($P < 0.05$) between WT and VEC mice).

have remained a debated issue, and our data provide decisive evidence that $\beta 2^*$ -nAChR receptors in neurons originating in the VTA have a crucial role.

Having established that re-expressed $\beta 2^*$ -nAChR receptors respond to nicotine *in vivo*, we investigated the potential role of endogenous acetylcholine in cognitive behaviours. The brain circuit containing the VTA, nucleus accumbens and ventral pallidum is involved in translation of environmental stimuli into adaptive responses, and is therefore implicated in exploratory and novelty-seeking behaviours^{19,20}. Exploratory behaviours contribute to the acquisition of environmental knowledge in rodents and other mammals²¹, and give rise to a variety of cognitive processes including spatial and non-spatial memory, spontaneous adaptation and strategy set-up²². Alterations in exploratory activity can be quantitatively analysed in rodents through detailed decompositions of the speed and location of animal trajectories in an open field²³. $\beta 2^{-/-}$ mice show normal motor function, anxiety processes and spatial and non-spatial memory systems²⁴ (see Supplementary Information). However, they also show modified spatio-temporal organization of displacements, with increased navigation and decreased exploratory behaviour compared with wild-type mice⁸.

We therefore tested whether the features of behaviour that distinguish wild-type from $\beta 2^{-/-}$ mice⁸ were rescued in VEC mice. Mice from the three experimental groups were placed in a circular open field and their trajectories measured using a video tracking system²³. WT and KO mice differed significantly in exploratory and navigation behaviours, confirming our published results on un.injected mice and demonstrating the absence of any lentiviral injection-related effects on these behaviours. However, VEC mice showed a selective restoration of exploratory behaviour, bringing this measure up to the level of WT mice, without significant modification of navigational behaviour (Fig. 4a). In addition, quantitative decomposition of trajectories (Fig. 4b and Supplementary Information) showed that the sequencing of movements in VEC mice is differentially restored, depending on the type of movements evaluated. In the central portion of the arena, VEC mice showed transitions between fast and slow movements that were similar to WT mice (Fig. 4c, right panel). This form of slow exploratory behaviour allows animals to gather spatial information external to the open field by making rearings and head movements²⁵. Conversely, VEC mice exhibited KO-like

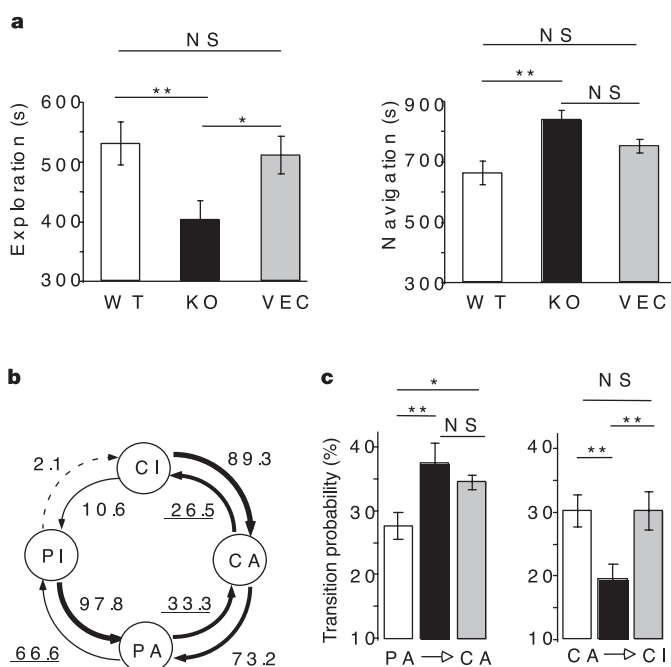


Figure 4 | VTA $\beta 2$ -subunit re-expression and behaviours in open fields. **a**, Time spent in exploration and navigation (mean $\pm$ s.e.m.). WT mice differ from KO mice in both exploration ($t = 0.0076$) and navigation behaviours ($t = 0.0004$). KO mice differ from VEC mice in exploration ($t = 0.02$) but not navigation behaviour ($t = 0.056$), whereas VEC mice are not (or are only marginally) different from WT mice (exploration, $t = 0.66$; navigation, $t = 0.049$). Single asterisk, $P < 0.01$; two asterisks, $P < 0.001$; NS, not significant. All comparisons were made using t -tests. **b**, Mean transition diagram. Frequency of shifting between movement types (see Methods) is indicated by arrow thickness, and conditional probabilities (in percentages) are shown. Underlined numbers are those transitions for which ANOVA revealed a significant group effect ($F_{2,27-P} = (3.81-0.034)$; $(3.80-0.035)$; $(4.34-0.023)$ and $(4.48-0.021)$ for PA-PI, PA-CA, CA-CI and CA-PA, respectively). **c**, Transition probability of PA-CA (left) and CA-CI (right) for WT, KO and VEC mice, shown as mean $\pm$ s.e.m. (single asterisk, $P < 0.05$; two asterisks, $P < 0.01$).

sequencing of fast navigational movements (Fig. 4c, left panel). The targeted expression of $\beta 2^*$ -nAChR in the VTA thus generates a dissociation between exploratory and navigational behaviour. Our results demonstrate that nAChR receptors in neurons originating in the VTA and/or their axonal projections suffice for the differential restoration of cognitive function, as measured by exploratory behaviour in our paradigm⁸. This behaviour therefore appears to be mediated by endogenous acetylcholine released either by projections from the pedunculo-pontine and latero-dorsal tegmental nucleus⁷, or by endogenous local acetylcholine acting on $\beta 2^*$ -nAChR receptors expressed on axonal projections from the VTA to the nucleus accumbens²⁶. It implicates dopamine or GABA as the neuromodulator for this executive function²⁰, as there is also an important projection of GABA-containing neurons to the prefrontal cortex²⁷.

Nicotine reinforcement has been postulated to mobilize multiple intricate networks^{28,29}, including ascending dopamine-, acetylcholine- and serotonin-mediated pathways together with glutamate- and GABA-containing neurons from multiple brain regions⁷. The present study reveals that restoration of $\beta 2^*$ -nAChR receptors, specifically in the VTA and its axonal projections, restores the self-administration of nicotine. The nAChR system in the VTA and its axonal projections is therefore a principal factor in nicotine reinforcement. Even more unexpected is the selective recovery of a complex cognitive behaviour as a result of cholinergic action, for which we provide a molecular ($\beta 2^*$ -nAChR receptors) and anatomical basis (neurons originating in the VTA).

This work further illustrates the efficiency of the lentivector technique *in vivo* in analyses of the neural bases of spontaneous cognitive behaviours and their regulation by endogenous neurotransmitters, and of a specific receptor species in the absence of external intervention. Molecular dissection of higher brain functions henceforth becomes accessible to *in vivo* investigation at the cellular and neuronal network level.

METHODS

Lentiviral expression vector. Vectors were based on pTRIPΔU3, modified as indicated in the Supplementary Information. The PGK-GFP lentivector is identical to the bi-cistronic version (that is, it also uses a phosphoglycerol kinase promoter) but lacks $\beta 2$ -IRES2 (the internal ribosomal entry sequence). Viral particles were generated as described in the Supplementary Information.

Stereotaxic procedure. Mice aged 6 to 12 weeks were anaesthetized using ketamine/xylazine in PBS. A mouse was introduced into a stereotaxic frame adapted for use with mice³⁰. Lentivector (2 μ l at 50 ng p24 protein per μ l) was injected bilaterally at: antero-posterior -3.4 mm, lateral ± 0.5 mm from bregma, and -4.4 mm from the surface. All procedures were carried out in accordance with European Commission directives 219/1990 and 220/1990, and approved by Animalerie centrale and Médecine du travail, Institut Pasteur. U. M. holds an Animal Surgery Authorization from the French Ministry of Agriculture.

[¹²⁵I]-epibatidine autoradiography. Experiments were performed as described previously¹¹ using 200 pM [¹²⁵I]-epibatidine (Perkin Elmer, specific activity 2200 Ci mmol⁻¹) in Tris 50 mM, pH 7.4 for 30 min.

***In vivo* electrophysiological recording of VTA dopaminergic neurons.** Single-unit extracellular recordings were performed in WT, KO and VEC mice (10 mice per group) as described¹³. Animals were anaesthetized with chloral hydrate (400 mg kg⁻¹ intraperitoneally). Spontaneously active dopamine neurons were identified on the basis of previously established electrophysiological criteria¹³. Intravenous injection of 0.5 mM nicotine tartrate into the saphenous vein (30 μ g kg⁻¹ free base in a final volume of ~ 10 μ l) was performed as described⁴. Firing frequency was quantified over 15-s periods, expressed as a percentage of average basal firing, and means were calculated within each group.

***In vivo* microdialysis.** *In vivo* microdialysis was carried out as described in the Supplementary Information, with a probe placed unilaterally in the shell of the nucleus accumbens, coordinates from bregma (in mm): anterior 1.34, lateral 0.7, ventral -5.4 . Probe placement was verified histologically. Analysis of variance (ANOVA) identified a significant effect of nicotine treatment ($F_{1,34} = 18.4$, $P < 0.001$) and genotype ($F_{2,34} = 5.6$, $P < 0.01$).

Intracranial self-administration. A Y-maze was used for morphine self-administration as described¹⁸. Mice were implanted unilaterally. In each experimental session, a guide cannula was used to insert an injection cannula, 1.5 mm beyond the tip of the guide cannula and into the VTA. The injection cannula contained a

solution of nicotine in Ringer's buffer set to pH 6.8. Mice were tested for nine consecutive days, with each daily session composed of 10 successive trials. The first two sessions (days 1 and 2) were used to habituate each mouse to the maze, and no drug was available. In subsequent sessions, interruption of photocell beams by entering the reinforced arm of the Y-maze triggered an intra-VTA injection of nicotine (100 ng nicotine tartrate in 50 nl), entering the non-reinforced arm produced no injection, and entering either arm terminated the trial. The number of times the mice entered the reinforced arm (that is, the number of self-administrations) and the mean latency to trigger injections (self-injection latency) were quantified. Neither wild-type nor $\beta 2^{-/-}$ mice showed preference for intra-VTA injections of vehicle (data not shown).

Quantitative analyses of locomotor behaviours. A video tracking system was used to break down trajectories into two components: navigation (large movements at a speed greater than or equal to 11.8 cm s⁻¹) and exploration (small movements at a speed less than or equal to 6.25 cm s⁻¹) (refs. 8, 23). The behavioural measures shown in Fig. 4a consisted of the time spent in both categories of displacements. ANOVA revealed significant group effects for exploration ($F_{2,27} = 4.825$, $P = 0.0162$) and navigation ($F_{2,27} = 8.24$, $P = 0.0016$). This system was also used to deconstruct movements into active or inactive (A and I) periods, and peripheral or central (P and C) positions within the open field by analyses of digitized video recordings taken at 25 frames s⁻¹. Active and inactive periods were defined by the instantaneous velocity of the mouse within the open field, and averaged using moving time-windows of 0.2 s. Peripheral and central positions for each mouse were defined by radial position of the mouse within two zones of the open field arena, termed peripheral (P) and central (C). Each frame was thus associated with two symbols; combining these symbols produced a four-symbol code (PI, CI, PA and CA) describing the state of the mouse at each time point during the experiment. Each individual experiment ($n = 10$ for each group) was then transformed into a sequence of 45,000 state symbols (30 min at 25 frames s⁻¹), which was subsequently reduced to a matrix of state transitions. This conditional matrix lists the probability of entering one state from another, with the sum of the conditional probabilities of entering the three other states from a given state equal to 1. A one-way ANOVA on transition probability values was performed to compare the effect of each experimental group. This analysis revealed significant main effects of experimental group (WT, KO and VEC) for the PA-to-CA and for the CA-to-CI transitions.

Statistical analyses. All data were analysed by repeated measures ANOVAs using StatView 5.0 (Abacus Concepts), followed by post-hoc Fisher's test for which the level of statistical significance was set at $P < 0.05$.

Received 25 January; accepted 28 April 2005.

- Peto, R. *et al.* Mortality from smoking worldwide. *Br. Med. Bull.* **52**, 12–21 (1996).
- Levin, E. D. & Rezvani, A. H. Nicotinic treatment for cognitive dysfunction. *Curr. Drug Targets CNS Neurol. Disord.* **1**, 423–431 (2002).
- Picciotto, M. R. *et al.* Abnormal avoidance learning in mice lacking functional high-affinity nicotine receptor in the brain. *Nature* **374**, 65–67 (1995).
- Picciotto, M. R. *et al.* Acetylcholine receptors containing the $\beta 2$ subunit are involved in the reinforcing properties of nicotine. *Nature* **391**, 173–177 (1998).
- Corrigan, P. J., Le Novère, N. & Changeux, J. P. Nicotinic receptors at the amino acid level. *Annu. Rev. Pharmacol. Toxicol.* **40**, 431–458 (2000).
- McGehee, D. S. & Role, L. W. Physiological diversity of nicotinic acetylcholine receptors expressed by vertebrate neurons. *Annu. Rev. Physiol.* **57**, 521–546 (1995).
- Laviolette, S. R. & van der Kooy, D. The neurobiology of nicotine addiction: bridging the gap from molecules to behaviour. *Nature Rev. Neurosci.* **5**, 55–65 (2004).
- Granon, S., Faure, P. & Changeux, J. P. Executive and social behaviors under nicotinic receptor regulation. *Proc. Natl Acad. Sci. USA* **100**, 9596–9601 (2003).
- Maskos, U., Kissa, K., St Clément, C. & Brulet, P. Retrograde trans-synaptic transfer of green fluorescent protein allows the genetic mapping of neuronal circuits in transgenic mice. *Proc. Natl Acad. Sci. USA* **99**, 10120–10125 (2002).
- Naldini, L. *et al.* *In vivo* gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science* **272**, 263–267 (1996).
- Zoli, M., Lena, C., Picciotto, M. R. & Changeux, J. P. Identification of four classes of brain nicotinic receptors using $\beta 2$ mutant mice. *J. Neurosci.* **18**, 4461–4472 (1998).
- Swanson, L. W. The projections of the ventral tegmental area and adjacent regions: a combined fluorescent retrograde tracer and immunofluorescence study in the rat. *Brain Res. Bull.* **9**, 321–353 (1982).
- Grace, A. A. & Bunney, B. S. Intracellular and extracellular electrophysiology of nigral dopaminergic neurons—I. Identification and characterization. *Neuroscience* **10**, 301–315 (1983).
- Grenhoff, J., Aston-Jones, G. & Svensson, T. H. Nicotinic effects on the firing pattern of midbrain dopamine neurons. *Acta Physiol. Scand.* **128**, 351–358 (1986).

15. Mansvelder, H. D., De Rover, M., McGehee, D. S. & Brussaard, A. B. Cholinergic modulation of dopaminergic reward areas: upstream and downstream targets of nicotine addiction. *Eur. J. Pharmacol.* **480**, 117–123 (2003).
16. Klink, R., de Kerchove d'Exaerde, A., Zoli, M. & Changeux, J. P. Molecular and physiological diversity of nicotinic acetylcholine receptors in the midbrain dopaminergic nuclei. *J. Neurosci.* **21**, 1452–1463 (2001).
17. Pontieri, F. E., Tanda, G., Orzi, F. & Di Chiara, G. Effects of nicotine on the nucleus accumbens and similarity to those of addictive drugs. *Nature* **382**, 255–257 (1996).
18. David, V., Durkin, T. P. & Cazala, P. Differential effects of the dopamine D2/D3 receptor antagonist sulpiride on self-administration of morphine into the ventral tegmental area or the nucleus accumbens. *Psychopharmacology (Berl.)* **160**, 307–317 (2002).
19. Bardo, M., Donohew, R. & Harrington, N. Psychobiology of novelty seeking and drug seeking behavior. *Behav. Brain Res.* **77**, 23–43 (1996).
20. Robbins, T. W. Chemical neuromodulation of frontal-executive functions in humans and other animals. *Exp. Brain Res.* **133**, 130–138 (2000).
21. Renner, M. J. Neglected aspects of exploratory behavior and investigatory behavior. *Psychobiology* **18**, 16–22 (1990).
22. Thinus-Blanc, C. *Animal Spatial Cognition: Behavioural and Neural Approaches* (World Scientific, Singapore, 1996).
23. Faure, P., Neumeister, H., Faber, D. & Korn, H. Symbolic analysis of swimming trajectories reveals scale invariance and provides a model for fish locomotion. *Fractals* **11**, 233–243 (2003).
24. Zoli, M., Picciotto, M. R., Ferrari, R., Cocchi, D. & Changeux, J. P. Increased neurodegeneration during ageing in mice lacking high-affinity nicotine receptors. *EMBO J.* **18**, 1235–1244 (1999).
25. Poucet, B. & Benhamou, S. The neuropsychology of spatial cognition in the rat. *Crit. Rev. Neurobiol.* **11**, 101–120 (1997).
26. Mesulam, M. M., Mufson, E. J., Wainer, B. H. & Levey, A. I. Central cholinergic pathways in the rat: an overview based on an alternative nomenclature (Ch1-Ch6). *Neuroscience* **10**, 1185–1201 (1983).
27. Carr, D. B. & Sesack, S. R. GABA-containing neurons in the rat ventral tegmental area project to the prefrontal cortex. *Synapse* **38**, 114–123 (2000).
28. Dehaene, S., Kerszberg, M. & Changeux, J. P. A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl Acad. Sci. USA* **95**, 14529–14534 (1998).
29. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
30. Cunningham, M. G. & McKay, R. D. A hypothermic miniaturized stereotaxic instrument for surgery in newborn rats. *J. Neurosci. Methods* **47**, 105–114 (1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This research was supported by the Institut Pasteur, Collège de France, Centre National de la Recherche Scientifique CNRS URA 2182 and UMR 5106, Association de Recherche sur le Cancer, European Commission Contracts 'NIDE' and 'Nicotine and Ageing', Mission Interministérielle de Lutte contre la Drogue et la Toxicomanie (MILDT), a National Research Service Award Fellowship from the NIH (to B. E. M.), a Scholarship from the Letten F. Saugstad Foundation (to M. B.), and an ATER from Collège de France (to A. E.). We are grateful to L. Prado for the electrophysiological characterization of expression plasmids, and to C. Reperant and M. Soudant for technical assistance. We would like to thank J.-P. Bourgeois, B. Gutkin, H. Korn, P.-M. Lledo and R. McKay for comments on the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.-P.C. (changeux@pasteur.fr).

LETTERS

ATP is a mediator of chemosensory transduction in the central nervous system

Alexander V. Gourine¹, Enrique Llaudet², Nicholas Dale² & K. Michael Spyer¹

Extracellular signalling by the purine nucleotide ATP has long been associated with sensory function^{1–8}. In the periphery, ATP mediates nociception^{3–5}, mechanosensitivity^{3,6}, thermal sensitivity⁷ and O₂ chemosensitivity⁸. These processes share a common mechanism that involves the release of ATP to excite afferent fibres via activation of ionotropic P2X and/or metabotropic P2Y receptors. Chemosensors located in the brainstem are crucial for the maintenance of physiological levels of blood gases through the regulation of breathing^{9–11}. Here we show that an increase in p_{CO_2} in the arterial blood triggers the immediate release of ATP from three chemosensitive regions located on the ventral surface of the medulla oblongata. Blockade of ATP receptors at these sites diminishes the chemosensory control of breathing, suggesting that ATP release constitutes a key step in central chemosensory transduction. These new data suggest that ATP, a phylogenetically ancient, unique and simple molecule, has been widely used in the evolution of afferent systems to mediate distinct forms of sensory transduction not only in the periphery but also within the central nervous system.

Every organism requires afferent (sensory) mechanisms to monitor its internal state and the environment in order to evoke life-preserving adaptive behavioural and physiological responses. Many peripheral sensory processes share a common mechanism that involves the release of ATP to excite afferent fibres via activation of ionotropic (P2X) and/or metabotropic (P2Y) cell-surface receptors^{1–8}. For nociception, cytosolic ATP released from damaged cells activates its receptors on C fibres to give the sensation of pain⁵. In the epithelia of the bladder and urethra, stretching causes the release of ATP and activation of P2X receptors on sensory afferents⁶. In the carotid body (the principal peripheral chemosensitive organ for regulation of breathing), O₂-sensitive glomus cells release ATP to activate P2X receptors on the afferent terminals of the carotid sinus nerve, which transmit information about arterial p_{O_2} levels to the respiratory centres of the brain⁸.

We hypothesized that ATP-mediated afferent transduction may also operate in the brain, where chemosensory mechanisms are vital for the regulation of osmotic balance, pH, p_{CO_2} and glucose concentration in the blood. Brainstem CO₂ chemoreception is ultimately responsible for adapting breathing to the needs of metabolism^{9–11}. Structures located on the ventral surface of the medulla oblongata function as primary central chemoreceptors that are essential to impart the effects of CO₂ on breathing^{9–11}. Although there is a consensus that changes in pH that follow changes in p_{CO_2} represent the appropriate stimulus for central chemoreceptors^{9–11}, their cellular identity and the signalling mechanisms involved remain largely unknown. Here we report that release of ATP from the classical⁹ brainstem chemosensitive areas mediates the effect of CO₂ on breathing, supporting the hypothesis that, similar to its role in the

periphery, ATP conveys afferent information within the central nervous system itself.

To detect ATP release in the brain we developed novel microelectrode biosensors to measure accurately changes in ATP concentration in real time^{12,13} (Fig. 1c, d). First, large (>1 mm in length; 100 μm in diameter) biosensors were placed in direct contact with a significant portion of the ventral surface CO₂ chemosensitive areas of the medulla (0.1–0.5 mm lateral from the pyramidal tracts and rostral from the XII nerve roots; Fig. 1c) of anaesthetized adult rats to determine whether ATP is released in response to an increase in the level of inspired CO₂. We detected an almost immediate release of ATP upon chemosensory stimulation (Fig. 1a, b) that had a mean peak amplitude of $3.8 \pm 0.9 \mu\text{M}$ ($n = 25$). The integrated response (measured by the large biosensors sampling from a significant portion of the ventral chemosensitive areas), expressed as the total amount of CO₂-evoked ATP released during 5 min of stimulation, was estimated to be $1,103 \pm 226 \mu\text{M}\cdot\text{s}$ ($n = 25$). To aid determination of the relative timing of ATP release and enhancement of ventilation, hypocapnic apnoea was induced by mechanical hyperventilation of some animals to reduce blood and brain levels of p_{CO_2} . Under these experimental conditions breathing was induced when the p_{CO_2} in arterial blood exceeded the apnoeic threshold. The increase in ATP release during hypercapnia occurred 19.5 ± 4.8 s before the induction of breathing (Fig. 1b). Similar delays between the CO₂-evoked release of ATP and enhancement in respiration were also recorded in normally breathing animals (eupnoea), confirming that under various baseline arterial p_{CO_2} conditions, ATP release from the ventral surface chemosensitive areas preceded the increase in breathing (see also Fig. 2a). The concentration of ATP decreased to baseline levels shortly after the termination of the CO₂ stimulus (Fig. 1a). The amount of ATP released in response to CO₂ in rats that had been sino-aortically denervated and in which the vagus nerves had been cut was similar to that for control animals (mean amplitude, $3.8 \pm 1.6 \mu\text{M}$, $n = 5$; total amount, $1,418 \pm 607 \mu\text{M}\cdot\text{s}$, $n = 5$). No release of ATP during hypercapnia was observed when the sensor was placed on the opposite (dorsal) surface of the brainstem. Thus, the release of ATP mirrors the CO₂ stimulus, is site specific—it occurs on the ventral but not the dorsal surface of the medulla—and does not require input from the peripheral chemoreceptors.

Miniature (125 or 250 μm in diameter) disk biosensors were then used to map the sites of CO₂-induced ATP production on the ventral surface of the medulla (Fig. 2a). We observed that during chemosensory stimulation, ATP was released on the ventral surface of the medulla in discrete locations closely corresponding to the classical CO₂ chemosensitive areas⁹ (Fig. 2a). Notably, the most rostral cluster of ATP-releasing sites (Fig. 2a) was in close proximity to a population of highly CO₂ chemosensitive neurons recently described in the retrotrapezoid nucleus¹¹.

¹Department of Physiology, Royal Free and University College London Medical School, Rowland Hill Street, London NW3 2PF, UK. ²Warwick Biosensor Group, Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK.

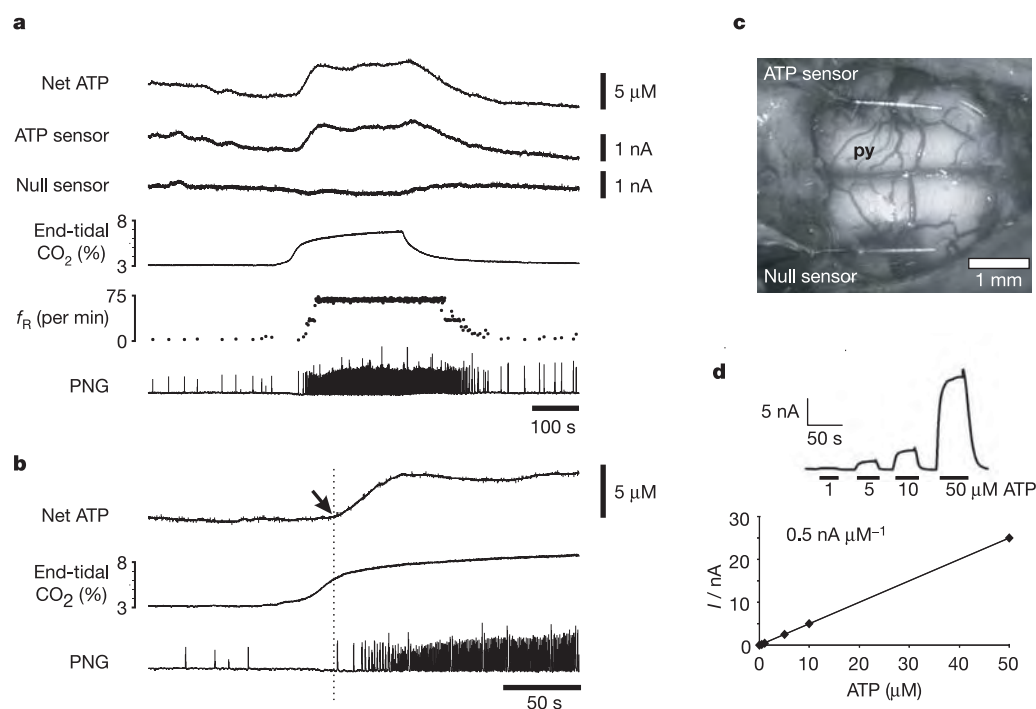


Figure 1 | Rapid CO_2 -induced release of ATP from the ventral surface of the medulla. **a**, Changes in respiratory activity (phrenic nerve discharge), ATP and null sensor currents in response to an increase in the level of inspired CO_2 in the rat. The 'Net ATP' trace represents the difference in signal between ATP and null sensors.

b, Expanded portion of **a** illustrating that ATP release precedes (arrow) respiratory activation. **c**, Placement of sensors on the ventral surface of the medulla. **d**, Current (I) calibration curve and sample responses of the ATP biosensor. f_R , respiratory frequency; PNG, phrenic neurogram (arbitrary units); py, pyramidal tract.

To dissect further the anatomical sites of CO_2 -induced ATP production, we isolated medullary chemosensitive structures *in vitro* by preparing horizontal slices of the rat medulla oblongata. CO_2 -induced acidification of the incubation media from pH 7.4 to 7.0 (an *in vitro* analogue of hypercapnia) evoked marked release of ATP (peak concentration $2.0 \pm 0.6 \mu\text{M}$, $n = 19$) from the most ventral slice, which contained surface chemosensitive areas (Fig. 2b, c). More dorsal slices in the sequence never exhibited significant ATP release in response to a decrease in pH (Fig. 2b, c). Thus, CO_2 -induced ATP release is highly localized to chemosensitive areas of the ventral surface of the medulla and occurs no more than $400 \mu\text{m}$ from this surface.

If ATP release from the ventral surface chemosensitive areas does indeed mediate CO_2 chemosensory transduction in the central nervous system, then blockade of ATP receptors in these areas should attenuate the CO_2 sensitivity of breathing. We therefore applied the ATP receptor antagonists pyridoxal-5'-phosphate-6-azophenyl-

2',4'-disulphonic acid (PPADS) and 2'-(or-3')-O-(trinitrophenyl)-adenosine 5'-triphosphate (TNP-ATP) to the medullary chemosensitive areas of anaesthetized adult rats (Fig. 3), and determined their respiratory responses to an increase in inspired CO_2 . Application of artificial cerebrospinal fluid (ACSF), PPADS or TNP-ATP had no effect on baseline respiratory activity during eupnoea, and did not induce breathing from hypocapnic apnoea.

The threshold level of end-tidal CO_2 required for inducing respiratory activity from hypocapnic apnoea after application of ACSF was $5.2 \pm 0.3\%$ ($n = 8$). It was raised significantly to $6.0 \pm 0.3\%$ ($n = 8$, $P < 0.05$) and $6.2 \pm 0.1\%$ ($n = 8$, $P < 0.05$) when PPADS was applied at concentrations of 10 and $100 \mu\text{M}$, respectively (Fig. 3a, b). Application of TNP-ATP had an equivalent effect (Fig. 3a, b). The peak amplitude of breathing during hypercapnia (end-tidal CO_2 8.0%; arterial p_{CO_2} 54 mm Hg, arterial pH 7.25) was also reduced by both ATP receptor antagonists (Fig. 3c). Thus, ATP receptor blockade on the ventral surface chemosensitive

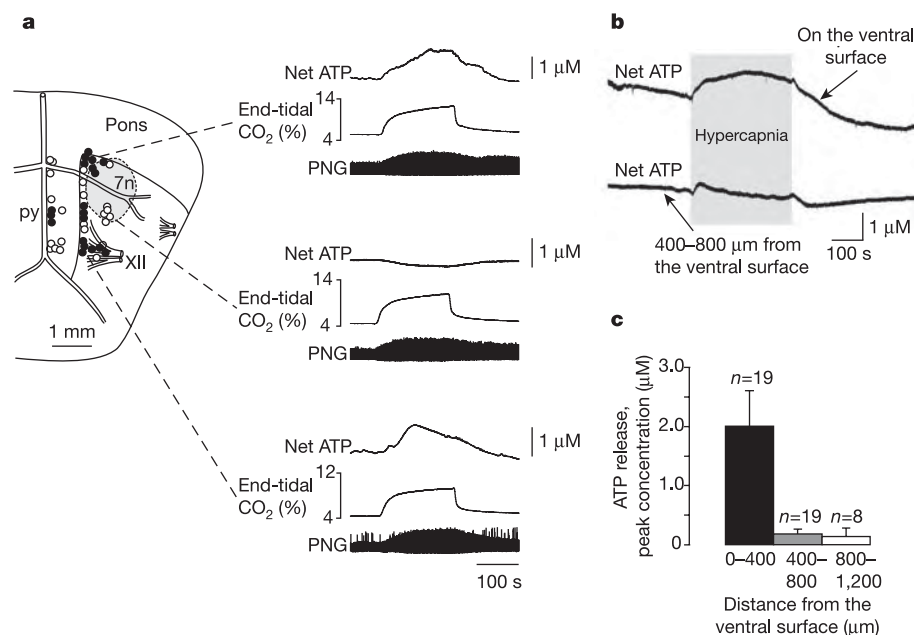


Figure 2 | CO_2 induces ATP release from the classical chemosensitive areas of the ventral medullary surface. **a**, Schematic drawing of the ventral aspect of the rat medulla oblongata showing sites that exhibited release of ATP in response to CO_2 (filled circles). No ATP release was detected in locations depicted by open circles. **b**, CO_2 -induced release of ATP from the ventral medullary surface *in vitro*. **c**, Summary data (means $\pm$ standard error) of peak CO_2/H^+ -induced release of ATP from horizontal slices of the medulla. ATP release occurs predominantly within $400 \mu\text{m}$ of the ventral surface. 7n, relative position of the facial nucleus; XII, hypoglossal nerve roots.

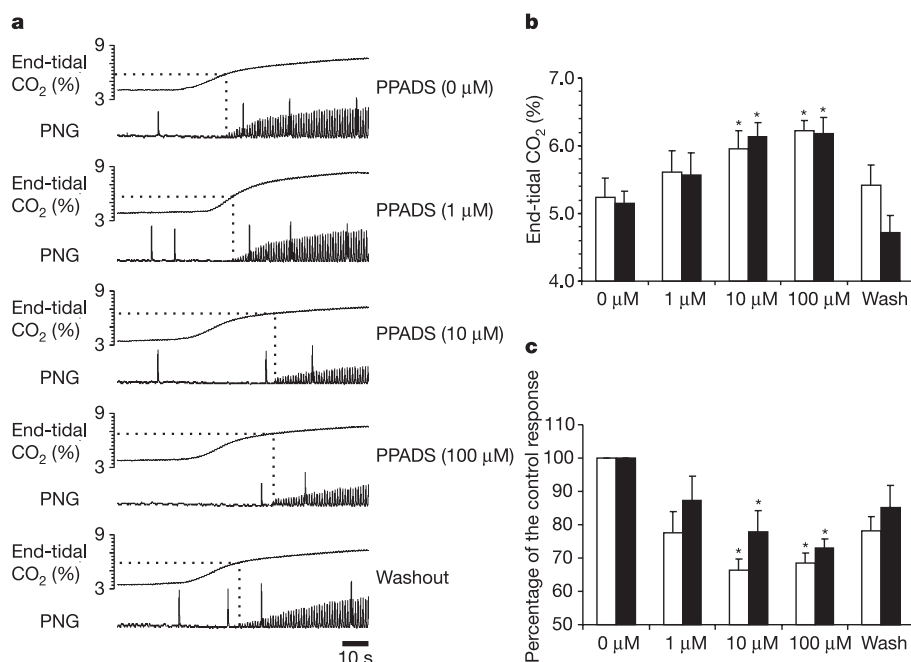


Figure 3 | ATP receptor blockade on the chemosensitive areas of the ventral medullary surface attenuates the effect of CO₂ on breathing in rats. **a**, Sequential recordings illustrating the effect of P2 antagonist PPADS on the threshold level of end-tidal CO₂ required to induce respiratory activity from hypocapnic apnoea. **b**, **c**, Summary data (means $\pm$ standard error) showing the increase in threshold levels of end-tidal CO₂ required to induce breathing from apnoea, and reduction in the CO₂-induced peak amplitude of the phrenic nerve bursts in the presence of PPADS (white columns; $n = 8$) or TNP-ATP (black columns; $n = 7$) on the ventral medullary surface. Asterisks indicate significant difference, $P < 0.05$.

areas reduces both the sensitivity and the gain of the respiratory response to rising levels of inspired CO₂. Although this reduction of chemosensitivity is incomplete, it reflects very closely the relative contribution of the surface chemoreceptors to the total respiratory response evoked by the action of CO₂ on all brainstem chemosensitive sites¹⁰.

We then applied ATP to the surface areas of the medulla that exhibited release of ATP in response to CO₂ to determine whether ATP can mimic the effect of CO₂ on breathing. ATP, when applied in this way, significantly increased the amplitude of respiratory activity (by around 25% ($P < 0.05$); time to peak of the response 27 ± 4 s; Fig. 4a, c). The effect of ATP was not affected by the adenosine receptor blocker 8-phenyltheophylline (8-PT), but was reduced by the ATP receptor antagonist suramin (Fig. 4a, d). Thus, application of exogenous ATP to the ventral surface chemosensitive areas of the

medulla mimics the effect of CO₂ on breathing. This is in contrast with the effect of ATP injected $\sim 500 \mu\text{m}$ dorsally, directly into the ventrolateral medulla, which causes suppression of respiration (Fig. 4b), demonstrating that the stimulatory effect of ATP on breathing is confined to the very surface region that releases ATP in response to CO₂. Application of the P2Y receptor agonist uridine 5'-triphosphate (UTP) to the ventral surface of the medulla had a delayed (time to peak of response 239 ± 41 s; $P < 0.01$) effect on the respiratory activity compared with ATP, but was ultimately equivalent in magnitude (Fig. 4a, c), suggesting that metabotropic P2Y receptors may, in part, mediate the action of ATP on breathing.

Three conjoining chemosensitive areas have been described on the ventral medullary surface⁹ just beneath the ventral respiratory column, which contains a network of respiratory neurons that are responsible for generation and shaping of the respiratory rhythm, as

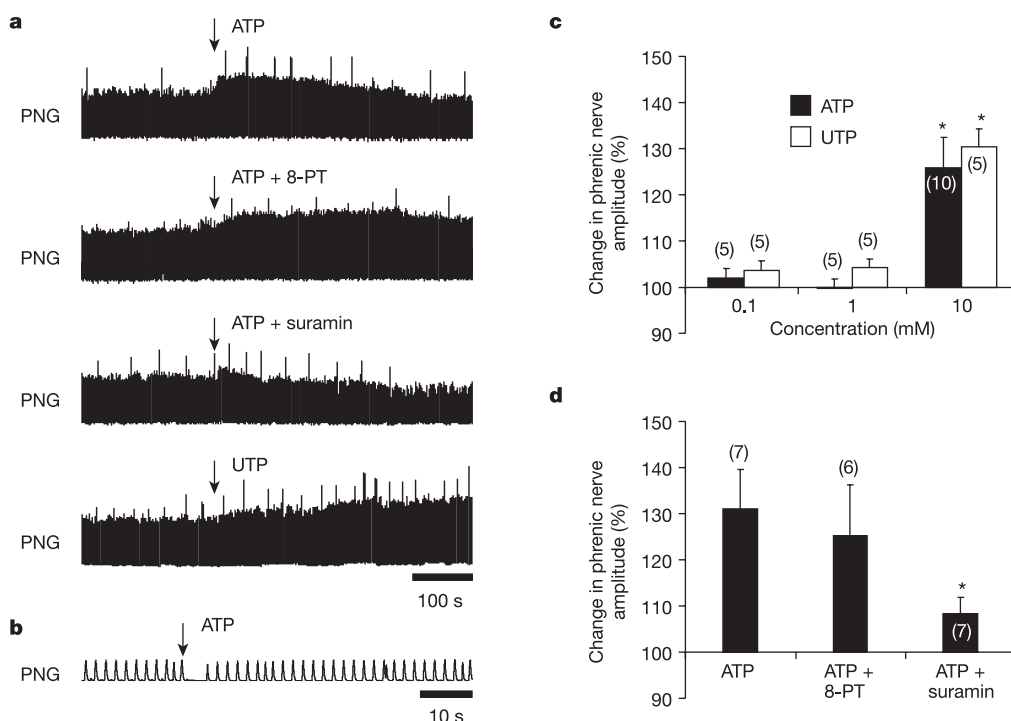


Figure 4 | ATP applied to the chemosensitive areas of the ventral medullary surface mimics the effect of CO₂ on breathing in rats. **a**, ATP and UTP (30- μl droplet of 10 mM) applied to the ventral medullary surface enhance respiratory activity. **b**, Microinjection of ATP (20 mM, 40 nl) into the rostral ventrolateral medulla inhibits breathing. **c**, Peak ATP- and UTP-induced (black and white columns, respectively) changes in the amplitude of the phrenic nerve discharge (means $\pm$ standard error). **d**, The effect of adenosine receptor blocker 8-PT (200 μM) or P2 antagonist suramin (20 mM) on ATP-induced increases in the amplitude of the phrenic nerve discharge (means $\pm$ standard error). Asterisks indicate significant difference, $P < 0.05$. Numbers in parentheses indicate sample size.

well as pre-motor neurons responsible for transmitting this rhythm to the respiratory muscles¹⁴. Our findings show that CO₂ induces immediate release of ATP from the chemosensitive structures located on the ventral surface of the medulla oblongata. When released, ATP can act on the distal dendrites of the ventral respiratory column neurons that project close to the ventral medullary surface¹⁵ to evoke adaptive enhancement in breathing. Some of these neurons express ATP receptors¹⁶.

The cellular sources and mechanisms underlying release of ATP in response to an increase in p_{CO_2} remain unknown. However, our data showing ATP release in the brain upon chemosensory stimulation will enable further detailed analysis of the chemotransduction process and identification of the primary chemosensitive cells. By extension it will also be of great interest and importance to establish whether ATP mediates other forms of chemosensitivity within the central nervous system; for example, osmo- and glucose sensitivity. Nevertheless, taking into account evidence indicating the key role for ATP in different peripheral afferent mechanisms and results of the present study, we propose a unifying hypothesis of central and peripheral sensory transduction that involves ATP as a common key mediator.

METHODS

See Supplementary Information for detailed Methods.

In vivo studies. Experiments were performed on 79 adult (age 10–12 weeks) male anaesthetized and artificially ventilated Sprague–Dawley rats, and carried out in accordance with the UK Animals (Scientific Procedures) Act, 1986. The CO₂ chemosensitive areas⁹ on the ventral surface of the medulla were exposed (Fig. 1c) to determine changes in ATP concentration during chemosensory stimulation. Activity of the phrenic nerve was recorded as an indicator of breathing activity. End-tidal levels of CO₂ were monitored online and kept at a designated level (either above or below the CO₂ apnoeic threshold) by altering respiratory volume and frequency. In all the experiments p_{O_2} in the arterial blood was kept at >100 mm Hg to ensure minimal drive from the peripheral chemoreceptors.

In vitro studies. Medullary chemosensitive structures were isolated *in vitro* by preparing horizontal slices of the ventral medulla oblongata of young adult rats (age 5 weeks; $n = 45$). Three slices (400- μm thickness) were cut, from which only the first one contained ventral surface chemosensitive areas. Recordings were made from the slices in a flow chamber at 33 °C.

ATP biosensors. The principal design and operation of the enzyme-based biosensors used in this paper have been described in detail elsewhere^{12,13}. The ATP sensors comprised two enzymes, glycerol kinase and glycerol-3-phosphate oxidase, entrapped within a matrix around a fine Pt wire of 100 μm in diameter and 1-mm long. The miniature sensors were formed from insulated Pt wire either 125 or 250 μm in diameter; the end of which had been cut to reveal a cross-section of bare Pt, which was then coated with the enzyme matrix. Electrochemical sensors can respond not only to the analyte of interest, but also to any electroactive species in the immediate environment. To check whether the sensors were responding to ATP, we used a dual recording configuration of ATP sensor placed upon one side of the medulla along with a null sensor that was placed in an equivalent position on the other side (Fig. 1c). The null sensor lacked the enzymes in the deposition layer and thus served as a control to determine whether any 'nonspecific' electroactive interferents were released and could confound the measurements.

Experimental protocols. For the measurement of CO₂-induced ATP release, ATP and null sensors were placed in direct contact with the designated areas of the ventral surface of the medulla oblongata. To determine the temporal relationship between changes in ATP concentration on the ventral surface and CO₂-evoked enhancement in the respiratory activity, hypocapnic apnoea was induced by mechanically hyperventilating some of the animals so that p_{CO_2} in the arterial blood and end-tidal levels of CO₂ were below the apnoeic threshold. Hypercapnia was then induced by titrating CO₂ into the respiratory mixture for a period of 3–5 min. In *in vitro* experiments, the analogue of hypercapnia was induced by perfusing the chamber with ACSF in which sufficient extra CO₂ had been bubbled to reduce the pH from its normal value of 7.4 to 7.0. ATP and null sensors were placed in equivalent positions on the ventral surface of the medulla.

To determine the effect of ATP receptor blockade on the CO₂-evoked enhancement in breathing, the P2 receptor antagonists PPADS or TNP-ATP were applied onto the ventral CO₂ chemosensitive areas and their effects on the respiratory response induced by an increase in inspired CO₂ were studied. To

determine accurately CO₂ chemosensitivity of breathing, hypocapnic apnoea was induced by mechanical hyperventilation of the animals. Then vehicle (ACSF), PPADS or TNP-ATP was applied to the exposed ventral chemosensitive areas and CO₂ was titrated into the respiratory mixture. The threshold level of end-tidal CO₂ required to induce respiratory activity from apnoea, and the peak amplitude of the phrenic nerve bursts during hypercapnia in the absence and presence of PPADS or TNP-ATP on the medullary surface, were determined.

ATP was applied on the ventral surface CO₂ chemosensitive areas of the medulla either alone or in combination with the adenosine receptor blocker 8-PT or ATP receptor antagonist suramin, and the effect of this treatment on breathing was determined. To investigate the role of metabotropic P2Y receptors, the effect of UTP was also determined. In these experiments animals were kept under normal ventilation, with arterial p_{CO_2} levels just above the apnoeic threshold.

Data analysis. Changes in ATP levels are presented as both raw data and means $\pm$ standard error of peak (in μM) and integral (in $\mu\text{M}\cdot\text{s}$) increases in concentrations. Integral increases in ATP concentrations were determined by measuring the area under the curve relative to a straight line joining the sensor current level before and after the response. The integral gives a better measure of the total production of analyte than the peak concentration. The data were compared using Student's paired *t*-test. A value of $P < 0.05$ was considered to be significant.

Received 3 March; accepted 25 April 2005.

- Chen, C. C. *et al.* A P2X purinoceptor expressed by a subset of sensory neurons. *Nature* **377**, 428–431 (1995).
- Lewis, C. *et al.* Coexpression of P2X₂ and P2X₃ receptor subunits can account for ATP-gated currents in sensory neurons. *Nature* **377**, 432–435 (1995).
- Cook, S. P., Vulchanova, L., Hargreaves, K. M., Elde, R. & McCleskey, E. W. Distinct ATP receptors on pain-sensing and stretch-sensing neurons. *Nature* **387**, 505–508 (1997).
- Cockayne, D. A. *et al.* Urinary bladder hyporeflexia and reduced pain-related behaviour in P2X₃-deficient mice. *Nature* **407**, 1011–1015 (2000).
- Cook, S. P. & McCleskey, E. W. Cell damage excites nociceptors through release of cytosolic ATP. *Pain* **95**, 41–47 (2002).
- Vlaskovska, M. *et al.* P2X₃ knock-out mice reveal a major sensory role for urothelially released ATP. *J. Neurosci.* **21**, 5670–5677 (2001).
- Souslova, V. *et al.* Warm-coding deficits and aberrant inflammatory pain in mice lacking P2X₃ receptors. *Nature* **407**, 1015–1017 (2000).
- Rong, W. *et al.* Pivotal role of nucleotide P2X₂ receptor subunit of the ATP-gated ion channel mediating ventilatory responses to hypoxia. *J. Neurosci.* **23**, 11315–11321 (2003).
- Loeschcke, H. H. Central chemosensitivity and the reaction theory. *J. Physiol. (Lond.)* **32**, 1–24 (1982).
- Feldman, J. L., Mitchell, G. S. & Nattie, E. E. Breathing: rhythmicity, plasticity, chemosensitivity. *Annu. Rev. Neurosci.* **26**, 239–266 (2003).
- Mulkey, D. K. *et al.* Respiratory control by ventral surface chemoreceptor neurons in rats. *Nature Neurosci.* **7**, 1360–1369 (2004).
- Llaudet, E., Botting, N. P., Crayston, J. A. & Dale, N. A three-enzyme microelectrode sensor for detecting purine release from central nervous system. *Biosens. Bioelectron.* **18**, 43–52 (2003).
- Llaudet, E., Hatz, S., Droniou, M. & Dale, N. Microelectrode biosensor for real-time measurement of ATP in biological tissue. *Anal. Chem.* **77**, 3267–3273 (2005).
- Richter, D. W. & Spyer, K. M. Studying rhythmogenesis of breathing: comparison of *in vivo* and *in vitro* models. *Trends Neurosci.* **24**, 464–472 (2001).
- Kawai, A., Ballantyne, D., Muckenhoff, K. & Scheid, P. Chemosensitive medullary neurones in the brainstem-spinal cord preparation of the neonatal rat. *J. Physiol. (Lond.)* **492**, 277–292 (1996).
- Gourine, A. V., Atkinson, L., Deuchars, J. & Spyer, K. M. Purinergic signalling in the medullary mechanisms of respiratory control in the rat: respiratory neurones express the P2X₂ receptor subunit. *J. Physiol. (Lond.)* **552**, 197–211 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. E. Droniou for help with the initial development of the ATP sensor. We also thank the Biotechnology and Biological Sciences Research Council (K.M.S., A.V.G.) and the Wellcome Trust (N.D.) for support.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to A.V.G. (a.gourine@medsch.ucl.ac.uk).

LETTERS

Angiotensin-converting enzyme 2 protects from severe acute lung failure

Yumiko Imai^{1*}, Keiji Kuba^{1*}, Shuan Rao², Yi Huan², Feng Guo², Bin Guan², Peng Yang², Renu Sarao¹, Teiji Wada¹, Howard Leong-Poi³, Michael A. Crackower⁴, Akiyoshi Fukamizu⁵, Chi-Chung Hui⁶, Lutz Hein⁷, Stefan Uhlig⁸, Arthur S. Slutsky⁹, Chengyu Jiang² & Josef M. Penninger¹

Acute respiratory distress syndrome (ARDS), the most severe form of acute lung injury, is a devastating clinical syndrome with a high mortality rate (30–60%) (refs 1–3). Predisposing factors for ARDS are diverse^{1,3} and include sepsis, aspiration, pneumonias and infections with the severe acute respiratory syndrome (SARS) coronavirus^{4,5}. At present, there are no effective drugs for improving the clinical outcome of ARDS^{1–3}. Angiotensin-converting enzyme (ACE) and ACE2 are homologues with different key functions in the renin–angiotensin system^{6–8}. ACE cleaves angiotensin I to generate angiotensin II, whereas ACE2 inactivates angiotensin II and is a negative regulator of the system. ACE2 has also recently been identified as a potential SARS virus receptor and is expressed in lungs^{9,10}. Here we report that ACE2 and the angiotensin II type 2 receptor (AT₂) protect mice from severe acute lung injury induced by acid aspiration or sepsis. However, other components of the renin–angiotensin system, including ACE, angiotensin II and the angiotensin II type 1a receptor (AT_{1a}), promote disease pathogenesis, induce lung oedemas and impair lung function. We show that mice deficient for *Ace2* show markedly improved disease, and also that recombinant ACE2 can protect mice from severe acute lung injury. Our data identify a critical function for ACE2 in acute lung injury, pointing to a possible therapy for a syndrome affecting millions of people worldwide every year.

The renin–angiotensin system has an important role in maintaining blood pressure homeostasis, as well as fluid and salt balance^{11–13}. ACE2 is a homologue of ACE, and functions as a negative regulator of the renin–angiotensin system^{6–8}. Although ACE2 is expressed in the lungs of humans¹⁰ and mice (see Supplementary Fig. 1a, b), nothing is known about its function in the lungs. However, mortality following SARS coronavirus infections approaches almost 10% owing to the development of ARDS^{14–16}. To elucidate the role of ACE2 in acute lung injury, we examined the effect of *Ace2* gene deficiency in mouse experimental models that mimic the common lung failure pathology observed in several human diseases, including sepsis, acid aspiration and pneumonias such as SARS and avian influenza A¹⁷.

Aspiration of gastric contents with a low pH is a frequent cause of acute lung injury/ARDS^{1–3}. Acid aspiration in wild-type mice, which mimics human acute lung injury^{18,19}, resulted in rapid impairment of lung functions assessed by increased lung elastance (a measure of the change in pressure achieved per unit change in volume, representing the stiffness of the lungs) (Fig. 1a), decreased blood oxygenation

(Fig. 1b) and the development of pulmonary oedema (Fig. 1c). Acid aspiration resulted in increased alveolar wall thickness, oedema, bleeding, inflammatory cell infiltrates and formation of hyaline membranes (Fig. 1d). Notably, acid-treated *Ace2* knockout mice⁸ showed significantly greater lung elastance compared with control wild-type mice, but there were no differences in lung elastance between saline-treated *Ace2* knockout and wild-type mice (Fig. 1a). Moreover, loss of *Ace2* resulted in worsened oxygenation (Fig. 1b), massive lung oedema (Fig. 1c), increased inflammatory cell infiltration and hyaline membrane formations (Fig. 1d) in response to acid aspiration. It should be noted that ACE2 protein expression is typically downregulated in wild-type mice following acid challenge (Fig. 1e).

Sepsis is the most common cause of acute lung injury/ARDS^{1–3}. We therefore examined the effect of *Ace2* gene deficiency on sepsis-induced acute lung injury using caecal ligation and perforation (CLP)²⁰. CLP causes lethal peritonitis and sepsis due to a polymicrobial infection that is accompanied by acute lung failure²⁰. Whereas all CLP-treated wild-type mice survived, only two out of ten CLP-treated *Ace2* knockout mice survived the 6 h experimental observation period (Fig. 2a). CLP resulted in lung failure defined by increased elastance (Fig. 2a), pulmonary oedema (Fig. 2b) and leukocyte accumulation (Fig. 2c) in wild-type mice. CLP-treated *Ace2* knockout mice had a marked worsening of lung functions (Fig. 2a), increased oedema (Fig. 2b) and leukocyte accumulation (Fig. 2c) compared with wild-type mice. In addition, *Ace2* knockout mice also developed markedly enhanced acute lung injury after endotoxin challenge¹⁸ (see Supplementary Fig. 2a–c). *Ace2* maps to the X chromosome, and it should be noted that loss of ACE2 expression resulted in equally severe acute lung injury phenotypes in male (*Ace2*^{−/Y}) and female (*Ace2*^{−/−}) mice. Our data from three different acute lung injury models show that loss of *Ace2* expression precipitates severe acute lung failure.

To test whether loss of ACE2 is essential for disease pathogenesis, we performed a rescue experiment using recombinant human ACE2 protein (rhuACE2) (see Supplementary Fig. 3a, b). Injection of rhuACE2 into acid-treated *Ace2* knockout mice decreased the degree of acute lung injury, as assessed by lung elastance (Fig. 2d) and pulmonary oedema formation (Fig. 2e). When we injected rhuACE2 protein into acid-treated wild-type mice, lung function (Fig. 2f) and oedema formation (see Supplementary Fig. 3c) were also rescued. In saline-treated wild-type or *Ace2* knockout mice, injections of

¹IMBA, Institute of Molecular Biotechnology of the Austrian Academy of Sciences, Vienna A-1030, Austria. ²National Laboratory of Medical Molecular Biology, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, China. ³Department of Cardiology, St. Michael's Hospital, Toronto, Ontario M5B 1W8, Canada. ⁴Department of Biochemistry and Molecular Biology, Merck Frosst Centre for Therapeutic Research, Montreal, Quebec H3R 4P8, Canada. ⁵Center for Tsukuba Advanced Research Alliance, University of Tsukuba, Tsukuba 305-8577, Japan. ⁶Program in Developmental Biology, The Hospital for Sick Children and Department of Molecular and Medical Genetics, University of Toronto, Toronto, Ontario M5S 1X8, Canada. ⁷Department of Pharmacology, University of Freiburg, Freiburg 79104, Germany. ⁸Division of Pulmonary Pharmacology, Research Center Borstel, Borstel 23845, Germany. ⁹Department of Medicine and Interdepartmental Division of Critical Care, University of Toronto, St. Michael's Hospital, Toronto, Ontario M5B 1W8, Canada.

*These authors contributed equally to this work.

rhuACE2 did not affect pulmonary functions (Fig. 2d–f). Catalytically inactive ACE2 protein (mut-rhuACE2) (see Supplementary Fig. 3a, b) did not rescue the severe lung phenotype in *Ace2* knockout mice (Fig. 2d, e) and had no effect on the severity of acute lung injury in wild-type animals (Fig. 2f and Supplementary Fig. 3c). These results show that the catalytic activity of ACE2 can directly protect lungs from acute lung injury.

ACE2 is a homologue of ACE, both of which are central enzymes in the renin–angiotensin system^{6–8}. ACE cleaves the decapeptide angiotensin (Ang)I into the octapeptide AngII (refs 11, 12). ACE2 cleaves a single residue from AngI to generate Ang1–9 (refs 6, 7), and a residue from AngII to generate Ang1–7 (ref. 6). In this way, ACE2 negatively regulates the renin–angiotensin system by inactivating AngII (Fig. 3a). Acid aspiration in wild-type mice resulted in marked downregulation of ACE2 protein, but ACE levels remained constant (Fig. 1e). Moreover, only catalytically active ACE2 improved the acute lung injury phenotype in mutant and wild-type mice (Fig. 2d–f). To clarify whether acute lung injury shifts the functional equilibrium between ACE and ACE2, we measured AngII levels in acid-treated and control mice. Acid aspiration markedly increased AngII levels in lungs (Fig. 3b) and plasma (see Supplementary Fig. 4a) of wild-type mice. We observed a further, significant increase in AngII levels in the lungs (Fig. 3b) and plasma (see Supplementary Fig. 4a) of acid-treated *Ace2* knockout mice. Thus, acute lung injury results in decreased ACE2 expression and increased production of AngII.

On the basis of these results we speculated that, in contrast to ACE2, ACE promotes disease pathogenesis through increased AngII production (Fig. 3a). Indeed, genetic inactivation of *Ace* on an *Ace2* wild-type or *Ace2* knockout background markedly decreased AngII levels in lung and plasma in our acid injury model (see Supplementary Fig. 4b, c). Moreover, treatment with rhuACE2 protein attenuated lung injury (Fig. 2d–f) and further reduced AngII levels in the lungs of acid-treated mice (Supplementary Fig. 4d). In contrast to *Ace2* knockout mice, *Ace*^{−/−} mice²¹ were partly protected against acute lung injury induced by acid-aspiration (Fig. 3c and Supplementary Fig. 5). These effects were dependent on gene dosage and were observed to a lesser extent in *Ace*^{+/-} mice. In addition,

inactivation of *Ace* on an *Ace2* knockout background rescued the severe lung failure (Fig. 3d), oedema formation (Fig. 3e) and histological changes (Fig. 3f) compared with *Ace2* knockout mice. Similarly, in endotoxin-induced acute lung injury, the severe lung impairments in *Ace2* knockout mice were reversed by additional *Ace* gene deficiency (see Supplementary Fig. 6). Thus, ACE promotes acute lung injury pathology and ACE2 alleviates it.

Both ACE and ACE2 are non-specific proteases that cleave additional substrates^{11,12}. Thus, although increased levels of AngII have been correlated with *Ace2* deficiency, it has not been shown that upregulation of the AngII pathway accounts for the observed phenotypes of *Ace2* knockout mice *in vivo*. The receptors for AngII in mice are angiotensin II type 1a (AT₁a) receptor²², the type 1b (AT₁b) receptor and the type 2 (AT₂) receptor²³. AT₁a and AT₂, but not AT₁b receptor expression is found in the lungs²⁵. We therefore explored which AngII receptor subtypes are responsible for ACE/ACE2 regulated acute lung injury, and whether AngII signalling through its receptors is responsible for ACE2-regulated lung pathology (Fig. 3a). Compared with wild-type mice, genetic loss of AT₁a receptor expression in *Agtr1a*^{−/−} mice²⁴ markedly improved lung function (Fig. 4a) and reduced oedema formation (see Supplementary Fig. 7a). In contrast, inactivation of the AT₂ receptor (*Agtr2*^{−/−})²⁵ aggravated acute lung injury (Fig. 4a and Supplementary Fig. 7a). AngII levels induced by acid aspiration in both *Agtr1a*^{−/−} and *Agtr2*^{−/−} mice were comparable to those in wild-type controls (not shown).

We next attempted to rescue acute lung injury in *Ace2* knockout mice using specific AT₁ and AT₂ receptor blockers. Pharmacological inhibition of AT₁ attenuated the severity of acid-induced lung injury in *Ace2* knockout mice (Fig. 4b and Supplementary Fig. 7b). Inhibition of AT₂ had no apparent effect on the acute lung injury phenotypes of *Ace2* knockout mice (Fig. 4b). These data show that the AT₁a and AT₂ receptors have opposite functions in controlling the severity of acute lung injury, and that actions of AngII through the AT₁a receptor have a causative role in acute lung failure.

Pulmonary oedema could arise from increased hydrostatic pressure (due to pulmonary vascular constriction) and/or enhanced

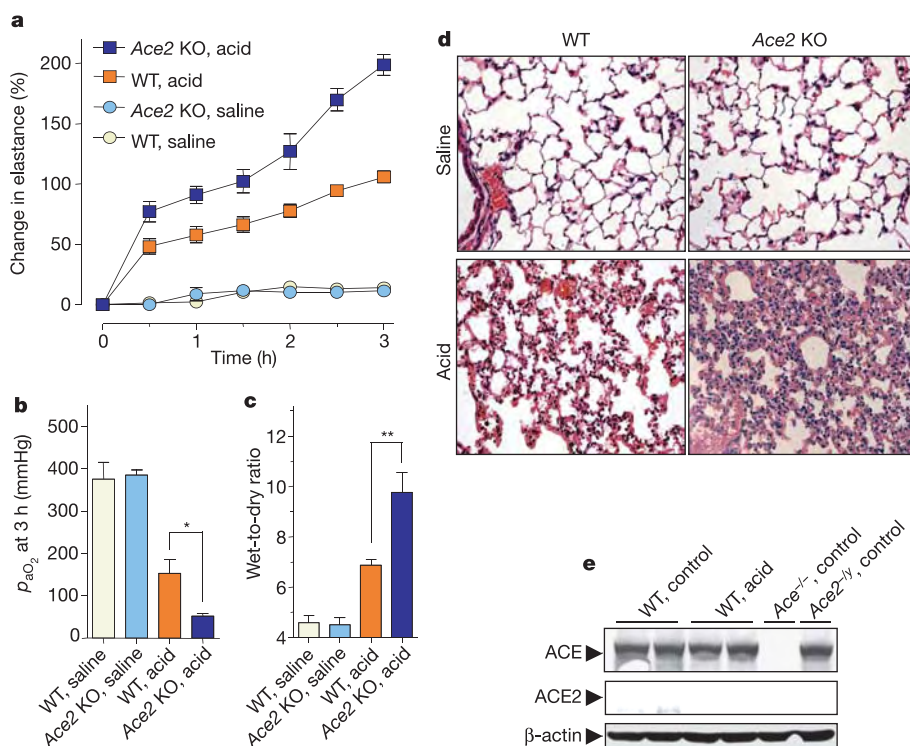


Figure 1 | Loss of ACE2 worsens acid aspiration-induced acute lung injury. **a**, Lung elastance after acid or saline treatment in wild type (WT) and *Ace2* knockout (*Ace2* KO) mice ($n = 10$ for acid-treated groups, $n = 6$ for saline-treated groups). $P < 0.05$ for the whole time course comparing acid-treated WT and *Ace2* knockout mice. **b**, Partial pressure of oxygen in arterial blood (p_{aO_2}) in acid-induced acute lung injury. **c**, Wet-to-dry weight ratios of lungs 3 h after acid injury. Single asterisk, $P < 0.05$; double asterisk, $P < 0.01$. **d**, Lung histopathology. Note the enhanced hyaline membrane formation, inflammatory cell infiltration and lung oedema in acid-treated *Ace2* knockout mice (H&E staining, $\times 200$). **e**, ACE and ACE2 protein expression in total lysates from control lungs and lungs 3 h after acid injury. Error bars indicate s.e.m.

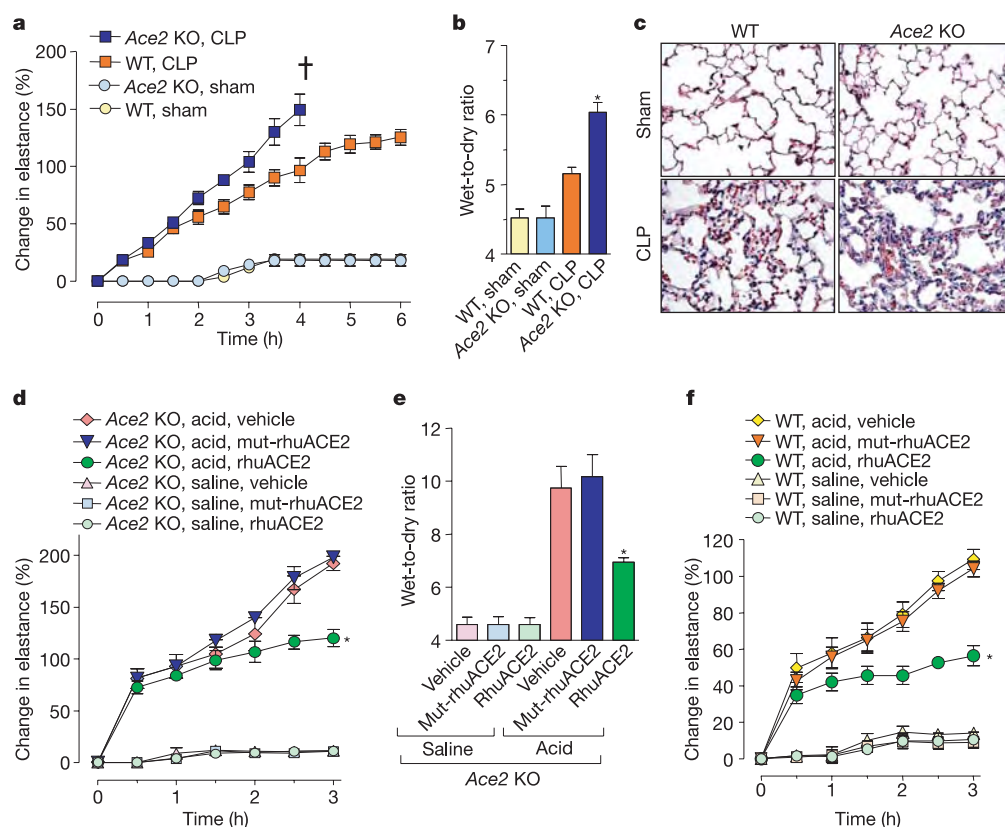


Figure 2 | ACE2 controls acute lung failure. **a**, Lung elastance after acute lung injury in WT and *Ace2* knockout (KO) mice induced by caecal ligation perforation (CLP). Eighteen hours after sham or CLP surgery, animals received mechanical ventilation for 6 h ($n = 10$ in CLP-treated groups, $n = 6$ in sham-treated groups). As 8/10 CLP-treated *Ace2* knockout mice died at 4–4.5 h, only data up to 4 h are shown. CLP-treated *Ace2* knockout mice had significantly higher elastance than CLP-treated WT mice ($P < 0.01$). **b**, **c**, Wet-to-dry weight ratios of lungs (**b**) and lung histopathology (**c**) in sham or CLP-treated WT and *Ace2* knockout mice determined after 4 h of ventilation. Asterisk denotes a significant difference ($P < 0.05$) between CLP-treated WT and *Ace2* knockout mice. Note the enhanced lung oedema and inflammatory infiltrates in *Ace2* knockout mice

(H&E staining, $\times 200$). **d**, **e**, Lung elastance (**d**) and wet-to-dry weight ratios (**e**) after acid or saline instillation of *Ace2* knockout mice injected intraperitoneally with recombinant human ACE2 protein (rhuACE2; 0.1 mg kg^{-1}), mutant rhuACE2 (mut-rhuACE2; 0.1 mg kg^{-1}) or vehicle ($n = 6$ per group). Asterisk denotes a significant difference ($P < 0.05$) comparing rhuACE2-treated *Ace2* knockout mice with mut-rhuACE2-treated and vehicle-treated *Ace2* knockout mice at 3 h. **f**, Lung elastance after acid instillation in WT mice treated with rhuACE2 protein (0.1 mg kg^{-1}), mut-rhuACE2 protein (0.1 mg kg^{-1}) or vehicle ($n = 6$ –8 per group). Asterisk denotes a significant difference ($P < 0.05$) between WT mice treated with rhuACE2 and mut-rhuACE2 or with vehicle at 3 h. Errors bars indicate s.e.m.

microvascular permeability²⁶. We first tested whether AngII can increase hydrostatic pressure using isolated, perfused murine lungs *ex vivo*²⁷. In this system, pulmonary perfusion pressures were comparable between wild-type and *Ace2* knockout mice under baseline control conditions (wild-type $3.0 \pm 1.9 \text{ cm H}_2\text{O}$, $n = 6$ versus *Ace2* knockout $1.8 \pm 1.6 \text{ cm H}_2\text{O}$, $n = 9$; mean $\pm$ s.e.m.), and these values were not changed by either acid-treatment or continuous perfusion of the bacterial endotoxin lipopolysaccharide (LPS). Pulmonary perfusion pressures generated by AngI or AngII injection into lungs of acid-instilled animals or into lungs perfused with LPS were also similar between wild-type and *Ace2* knockout mice (see Supplementary Fig. 8a, b). Moreover, fractional shortening using echocardiography (an indicator of left ventricular systolic function) and mean arterial pressures were comparable between *Ace2* knockout and wild-type mice during the experimental period (see Supplementary Fig. 9a, b). Thus, the severe lung oedemas in *Ace2* knockout mice do not seem to be secondary to systemic haemodynamic alterations.

As enhanced pulmonary vascular permeability is a hallmark of acute lung injury/ARDS in humans², we examined whether loss of *Ace2* results in increased vascular permeability using Evans Blue dye injections as an *in vivo* indicator of albumin leakage²⁸. Acid aspiration increased vascular permeability in wild-type mice. In *Ace2*

knockout mice, pulmonary Evans Blue accumulation was greatly increased after acid aspiration (Fig. 4c, d). These results were confirmed using fluorescein isothiocyanate (FITC)-conjugated dextran (40 kDa) as another marker to assess vascular leakage of macromolecules (data not shown). Vascular permeability was significantly attenuated in the lungs of *Agtr1a*^{-/-} mice (Fig. 4e). We suggest that loss of ACE2 expression in acute lung injury leads to leaky pulmonary blood vessels through AT₁ receptor stimulation. However, hydrostatic oedemas cannot be excluded, and the effects of local AngII production on lung blood vessels require further investigation^{27,29}.

ARDS is the most severe form of a wide spectrum of pathological processes designated as acute lung injury². ARDS is characterized by pulmonary oedema due to increased vascular permeability, the accumulation of inflammatory cells and severe hypoxia². Predisposing factors for ARDS include sepsis, aspiration and pneumonias (including infections with SARS coronavirus^{1–5} or avian influenza viruses¹⁷). Our data show that acute lung injury results in a marked downregulation of ACE2, a key enzyme involved in the regulation of the renin–angiotensin system.

It has been previously shown that an insertion/deletion ACE polymorphism that affects ACE activity is associated with ARDS susceptibility and outcome³⁰. Our data provide a mechanistic

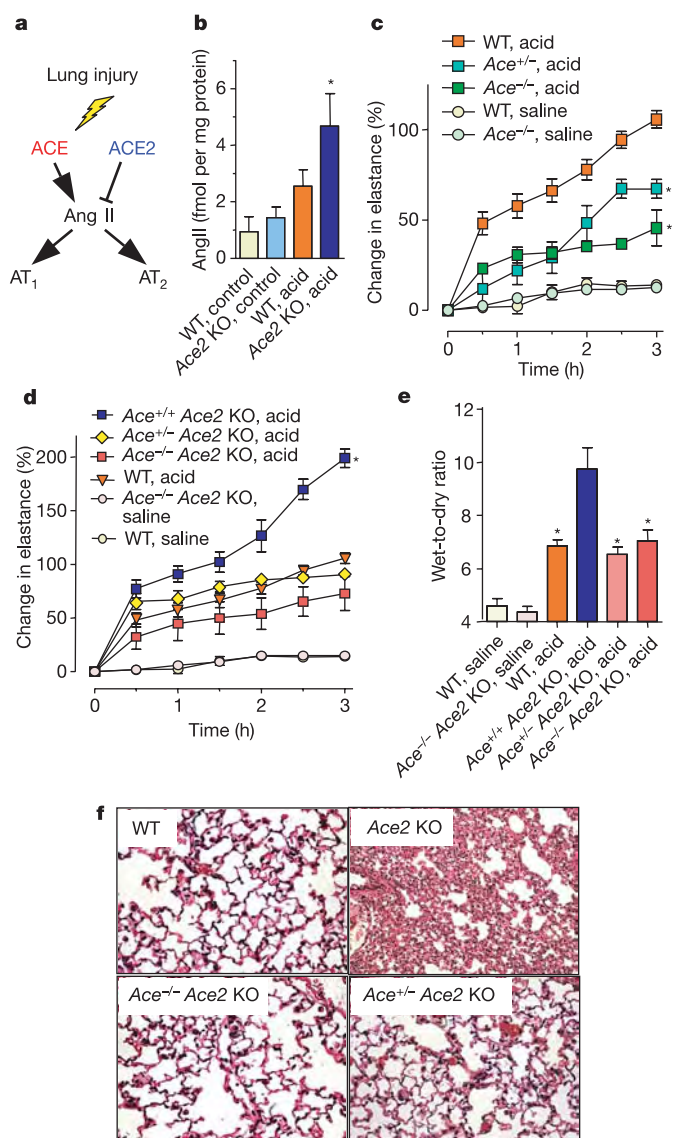


Figure 3 | ACE deficiency reduces the severity of acute lung injury. **a**, Schematic diagram of the renin-angiotensin system. **b**, Lung levels of AngII in control and acid-treated WT and Ace2 knockout (KO) mice determined at 3 h by enzyme immunoassay ($n = 3-5$ per group). Asterisk denotes a significant difference ($P < 0.05$) between acid-treated WT and Ace2 knockout mice. **c**, Lung elastance after acid instillation in Ace^{+/+} (WT), Ace^{+/+} and Ace^{-/-} mice ($n = 4-6$ mice per group). Asterisk denotes a significant difference ($P < 0.05$) comparing Ace^{+/+} with Ace^{+/+} and Ace^{-/-} mice at 3 h. **d**, **e**, Lung elastance (**d**) and wet-to-dry lung weight ratios (**e**) in acid- or saline-treated Ace^{+/+} Ace2 KO, Ace^{+/+} Ace2 KO, Ace^{-/-} Ace2 KO and WT mice ($n = 5$ per group). Asterisk denotes a significant difference ($P < 0.05$) comparing Ace2 KO with WT, Ace^{+/+} Ace2 KO or Ace^{-/-} Ace2 KO mice 3 h after acid-treatment. **f**, Lung histopathology. Severe lung interstitial oedema and leukocyte infiltration in Ace2 KO mice are attenuated by homozygous (Ace^{-/-}) or heterozygous (Ace^{+/-}) mutations of Ace (H&E staining, ×200). Error bars indicate s.e.m.

explanation for these clinical findings and indicate that, in the pathogenesis of acute lung injury, AngII is upregulated by ACE and drives severe lung failure through the AT₁ receptor. On the other hand, ACE2 and the AT₂ receptor protect against lung injury. Exogenous recombinant human ACE2 attenuates acute lung failure in Ace2 knockout as well as in wild-type mice. This combination of genetic, pharmacological and protein rescue experiments defines a new and critical role for the renin-angiotensin system in the pathogenesis of acute lung injury, and show that ACE2 is a key

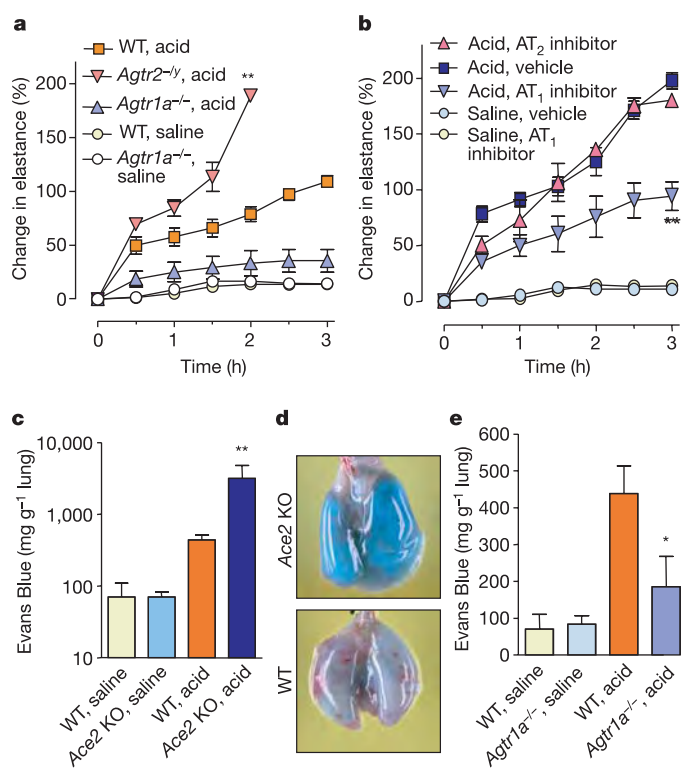


Figure 4 | The AngII receptor AT₁a controls acute lung injury severity and pulmonary vascular permeability. **a**, Lung elastance measurements in Agtr1a^{-/-} mice, Agtr2^{-/-} mice and WT mice after acid aspiration ($n = 4-6$ per group). All acid-treated Agtr2^{-/-} mice died after 2 h. There is a significant difference ($P < 0.01$) between acid-treated WT and acid-treated Agtr1a^{-/-} mice over the whole time course. Double asterisk denotes a significant difference ($P < 0.01$) between WT and Agtr2^{-/-} mice at 2 h. **b**, Lung elastance measurements in Ace2 knockout mice treated with vehicle or inhibitors to AT₁ (Losartan, 15 mg kg⁻¹) or AT₂ (PD123.319, 15 mg kg⁻¹) after acid or saline instillation (see Methods, $n = 4-6$ per group). Double asterisk denotes a significant difference ($P < 0.01$) comparing Ace2 knockout mice treated with AT₁ inhibitor with vehicle or AT₂ inhibitor treatment at 3 h. **c**, Pulmonary vascular permeability as determined by intravenous injection of Evans Blue. Extravascular Evans Blue in lungs was measured in WT and Ace2 knockout mice 3 h after acid injury ($n = 5$ per group). Double asterisk denotes a significant difference ($P < 0.01$) between acid-treated WT and Ace2 knockout mice. **d**, Representative images of Evans Blue-injected lungs of WT and Ace2 knockout mice 3 h after acid aspiration. **e**, Extravascular Evans Blue in lungs of WT and Agtr1a^{-/-} mice 3 h after acid injury ($n = 5$ per group). Asterisk denotes a significant difference ($P < 0.05$) between acid-treated WT and Agtr1a^{-/-} mice at 3 h. Error bars indicate s.e.m.

molecule involved in the development and progression of acute lung failure.

METHODS

For detailed methods please refer to the Supplementary Information.

Animals. Ace2, Ace, Agtr1a and Agtr2 mutant mice have previously been described^{2,4,5,6}. Sex-, age-, and background-matched mice were used as controls. Basal lung functions and lung structure were comparable among all the mice tested. Mice were handled in accordance with institutional guidelines.

Experimental murine models of acute lung injury. For acid aspiration-induced acute lung injury, anaesthetized mice were intratracheally instilled with HCl (pH 1.5; 2 ml kg⁻¹) and ventilated for 3 h (refs 18, 19). To study sepsis-induced acute lung injury, we performed caecal ligation perforation (CLP)²⁰. Sham-operated mice underwent the same procedure without ligation and puncture of the caecum. Eighteen hours after sham/CLP surgery, animals were subjected to mechanical ventilation for up to 6 h. For endotoxin-induced acute lung injury, anaesthetized mice received LPS and zymosan intratracheally immediately after starting mechanical ventilation and 1 h later, respectively¹⁸. In all acute lung

injury models, total positive end expiratory pressure ($PEEP_t$) and plateau pressure (P_{plat}) were measured at the end of expiratory and inspiratory occlusion, respectively. Elastance was calculated as $(P_{plat} - PEEP_t)$ divided by tidal volume (V_T) every 30 min during the ventilation periods.

Blood oxygenation, pulmonary oedema, pulmonary vascular permeability and histology. Blood samples were obtained from the left heart ventricle and partial pressure of oxygen in arterial blood (p_{aO_2}) was measured. To assess pulmonary oedemas, the lung wet-to-dry weight ratios were calculated. Pulmonary vascular permeability was assessed by measuring the pulmonary extravasation of Evans Blue. For histological analysis, 5- μ m thick sections were cut and stained with haematoxylin and eosin (H&E).

Recombinant ACE2 and AT₁/AT₂ receptor inhibitors. Thirty minutes before acid instillation, mice received intraperitoneal injections of recombinant human ACE2 (rhuACE2) protein (0.1 mg kg^{-1}) (R&D Systems or our own rhuACE2 preparation), catalytically inactive (H374N, H378N)¹⁰ mutant recombinant human ACE2 (mut-rhuACE2) or vehicle (0.1% BSA/PBS). All animals were then ventilated for 3 h. RhuACE2 protein and mut-rhuACE2 protein were purified from transfected CHO cells by affinity chromatography. The catalytic activities of purified recombinant ACE2 proteins were measured using the fluorogenic peptide Substrate VI (R&D Systems). Mut-rhuACE2-Fc showed >95% loss of catalytic activity (see Supplementary Fig. 3a). For inhibitor studies, mice received intraperitoneal injections of the AT₁ inhibitor Losartan (15 mg kg^{-1}), the AT₂ inhibitor PD123.319 (15 mg kg^{-1}) or control vehicle 30 min before surgical procedures.

Angiotensin II peptide levels and western blotting. AngII peptide levels were measured as described⁸. For western blotting, rabbit polyclonal anti-ACE2 antibody⁸ and rabbit polyclonal anti-mouse ACE antibody (R&D Systems) were used.

Statistical analyses. All data are shown as mean $\pm$ s.e.m.. Measurements at single time points were analysed by analysis of variance (ANOVA). Time courses were analysed by repeated measurements ANOVA with Bonferroni post-tests.

Received 11 February; accepted 29 April 2005.

- Hudson, L. D., Milberg, J. A., Anardi, D. & Maunder, R. J. Clinical risks for development of the acute respiratory distress syndrome. *Am. J. Respir. Crit. Care Med.* **151**, 293–301 (1995).
- Ware, L. B. & Matthay, M. A. The acute respiratory distress syndrome. *N. Engl. J. Med.* **342**, 1334–1349 (2000).
- Vincent, J. L., Sakr, Y. & Ranieri, V. M. Epidemiology and outcome of acute respiratory failure in intensive care unit patients. *Crit. Care Med.* **31** (suppl.), S296–S299 (2003).
- Ksiazek, T. G. *et al.* A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**, 1953–1966 (2003).
- Drosten, C. *et al.* Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**, 1967–1976 (2003).
- Donoghue, M. *et al.* A novel angiotensin-converting enzyme-related carboxypeptidase (ACE2) converts angiotensin I to angiotensin 1–9. *Circ. Res.* **87**, E1–E9 (2000).
- Tipnis, S. R. *et al.* A human homolog of angiotensin-converting enzyme. Cloning and functional expression as a captopril-insensitive carboxypeptidase. *J. Biol. Chem.* **275**, 33238–33243 (2000).
- Crackower, M. A. *et al.* Angiotensin-converting enzyme 2 is an essential regulator of heart function. *Nature* **417**, 822–828 (2002).
- Li, W. *et al.* Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **426**, 450–454 (2003).
- Hamming, I. *et al.* Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. *J. Pathol.* **203**, 631–637 (2004).
- Skeggs, L. T., Dorer, F. E., Levine, M., Lentz, K. E. & Kahn, J. R. The biochemistry of the renin-angiotensin system. *Adv. Exp. Med. Biol.* **130**, 1–27 (1980).
- Corvol, P., Williams, T. A. & Soubrier, F. Peptidyl dipeptidase A: angiotensin I-converting enzyme. *Methods Enzymol.* **248**, 283–305 (1995).
- Boehm, M. & Nabel, E. G. Angiotensin-converting enzyme 2—a new cardiac regulator. *N. Engl. J. Med.* **347**, 1795–1797 (2002).
- Tsang, K. W. *et al.* A cluster of cases of severe acute respiratory syndrome in Hong Kong. *N. Engl. J. Med.* **348**, 1977–1985 (2003).
- Lee, N. *et al.* A major outbreak of severe acute respiratory syndrome in Hong Kong. *N. Engl. J. Med.* **348**, 1986–1994 (2003).
- Poutanen, S. M. *et al.* Identification of severe acute respiratory syndrome in Canada. *N. Engl. J. Med.* **348**, 1995–2005 (2003).
- Tran, T. H. *et al.* Avian influenza A (H5N1) in 10 patients in Vietnam. *N. Engl. J. Med.* **350**, 1179–1188 (2004).
- Nagase, T. *et al.* Acute lung injury by sepsis and acid aspiration: a key role for cytosolic phospholipase A2. *Nature Immunol.* **1**, 42–46 (2000).
- Imai, Y. *et al.* Injurious mechanical ventilation and end-organ epithelial cell apoptosis and organ dysfunction in an experimental model of acute respiratory distress syndrome. *J. Am. Med. Assoc.* **289**, 2104–2112 (2003).
- Martin, E. L. *et al.* Negative impact of tissue inhibitor of metalloproteinase-3 null mutation on lung structure and function in response to sepsis. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **285**, L1222–L1232 (2003).
- Krege, J. H. *et al.* Male–female differences in fertility and blood pressure in ACE-deficient mice. *Nature* **375**, 146–148 (1995).
- Inagami, T. *et al.* Cloning, expression and regulation of angiotensin II receptors. *Adv. Exp. Med. Biol.* **377**, 311–317 (1995).
- Gasc, J. M., Shanmugam, S., Sibony, M. & Corvol, P. Tissue-specific expression of type 1 angiotensin II receptor subtypes. An *in situ* hybridization study. *Hypertension* **24**, 531–537 (1994).
- Sugaya, T. *et al.* Angiotensin II type 1a receptor-deficient mice with hypotension and hyperreninemia. *J. Biol. Chem.* **270**, 18719–18722 (1995).
- Hein, L., Barsh, G. S., Pratt, R. E., Dzau, V. J. & Kobilka, B. K. Behavioural and cardiovascular effects of disrupting the angiotensin II type-2 receptor in mice. *Nature* **377**, 744–747 (1995).
- Plante, G. E., Chakir, M., Ettaouil, K., Lehoux, S. & Sirois, P. Consequences of alteration in capillary permeability. *Can. J. Physiol. Pharmacol.* **74**, 824–833 (1996).
- Roy, B. J., Pitts, V. H. & Townsley, M. I. Pulmonary vascular response to angiotensin II in canine pacing-induced heart failure. *Am. J. Physiol.* **271**, H222–H227 (1996).
- Goggel, R. *et al.* PAF-mediated pulmonary edema: a new role for acid sphingomyelinase and ceramide. *Nature Med.* **10**, 155–160 (2004).
- Hansen, T. N., Le Blanc, A. L. & Gest, A. L. Hypoxia and angiotensin II infusion redistribute lung blood flow in lambs. *J. Appl. Physiol.* **58**, 812–818 (1985).
- Marshall, R. P. *et al.* Angiotensin converting enzyme insertion/deletion polymorphism is associated with susceptibility and outcome in acute respiratory distress syndrome. *Am. J. Respir. Crit. Care Med.* **166**, 646–650 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Chappell, C. Richardson, B. Seed, U. Eriksson, J. Ishida and all members of our laboratory for discussions and reagents. This work is supported by the Institute for Molecular Biotechnology of the Austrian Academy of Sciences (IMBA) and the Jubilaeumsfonds of the Austrian National Bank. This work is in part supported by the Canadian Institutes of Health Research (CIHR) and the Canada Foundation for Innovation (CFI). K.K. is supported by a Marie Curie Fellowship from the EU. C.J. is supported a Beijing Committee of Science and Technology grant and the Natural Science Foundation of China. L.H. is supported by the Deutsche Forschungsgemeinschaft.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to J.M.P. (Josef.penninger@imba.oew.ac.at).

Integrative genomic analyses identify *MITF* as a lineage survival oncogene amplified in malignant melanoma

Levi A. Garraway^{1,3}, Hans R. Widlund¹, Mark A. Rubin^{2,3}, Gad Getz⁵, Aaron J. Berger⁶, Sridhar Ramaswamy^{5,7}, Rameen Beroukhi^{1,3}, Danny A. Milner^{2,3}, Scott R. Granter², Jinyan Du^{1,5}, Charles Lee^{2,3}, Stephan N. Wagner⁸, Cheng Li^{1,4}, Todd R. Golub^{1,3,5}, David L. Rimm⁶, Matthew L. Meyerson^{1,2,5}, David E. Fisher^{1,3} & William R. Sellers^{1,2,3,5}

Systematic analyses of cancer genomes promise to unveil patterns of genetic alterations linked to the genesis and spread of human cancers. High-density single-nucleotide polymorphism (SNP) arrays enable detailed and genome-wide identification of both loss-of-heterozygosity events and copy-number alterations in cancer^{1–5}. Here, by integrating SNP array-based genetic maps with gene expression signatures derived from NCI60 cell lines, we identified the melanocyte master regulator *MITF* (microphthalmia-associated transcription factor) as the target of a novel melanoma amplification. We found that *MITF* amplification was more prevalent in metastatic disease and correlated with decreased overall patient survival. *BRAF* mutation and p16 inactivation accompanied *MITF* amplification in melanoma cell lines. Ectopic *MITF* expression in conjunction with the *BRAF*(V600E) mutant transformed primary human melanocytes, and thus *MITF* can function as a melanoma oncogene. Reduction of *MITF* activity sensitizes melanoma cells to chemotherapeutic agents. Targeting *MITF* in combination with *BRAF* or cyclin-dependent kinase inhibitors may offer a rational therapeutic avenue into melanoma, a highly chemotherapy-resistant neoplasm. Together, these data suggest that *MITF* represents a distinct class of 'lineage survival' or 'lineage addiction' oncogenes required for both tissue-specific cancer development and tumour progression.

To begin to organize human cancers on the basis of large-scale chromosomal alterations, we first evaluated the genomes of NCI60 cell lines⁶ using pre-release 100K SNP arrays (Affymetrix). These arrays interrogate over 124,000 SNP alleles spaced with a median intermarker distance of 8.5 kilobases (kb). The NCI60 cell lines represent tumours from nine different tissue types and are annotated by multiple large-scale data sets^{7–10}. Thus, these cell lines offer a platform for testing integrated and orthogonal analytic approaches (Fig. 1a). The complete NCI60 SNP array data are available at http://dtp.nci.nih.gov/mtargets/mt_index.html and at <http://www.ncbi.nlm.nih.gov/geo> (accession number GSE2520).

To determine whether patterns of copy number alterations identified distinct genetic subgroups, hierarchical clustering¹¹ was used to organize 58 NCI60 cell lines based on copy-number alterations defined by SNP array analysis (see Methods). The resulting dendrogram contained subclusters in which the samples segregated largely according to tissue of origin (Fig. 1b, Supplementary Fig. 1). One

such group consisted of lung cancer lines, another mainly of colon tumour lines, and another of cell lines derived from malignant melanoma (Fig. 1b). For each of these cell line clusters, there were also associated SNP clusters derived from contiguous chromosomal regions (Fig. 1b and Supplementary Fig. 1).

The tissue-based organization of the samples raised the possibility that the associated chromosomal aberrations might harbour lineage-specific cancer genes. To investigate this, we focused on a region of copy gain on chromosome 3 at 3p13–3p14 that defined the melanoma subcluster (Fig. 1b,c). Here, supervised analysis¹² (see Supplemental Methods) was performed using available NCI60 gene expression data, looking for gene expression correlates of the class distinction '3p amplified' (six melanoma cell lines) versus '3p non-amplified' (remaining NCI60 lines; Fig. 1c). 583 transcripts demonstrated significantly increased expression in the 3p-amplified class after Bonferroni correction, with $P < 10^{-6}$ by t -statistic. This expression signature was influenced significantly by lineage-related differences in transcript profiles between melanomas and the other NCI60 tissue types, as noted by others⁷. Nonetheless, only one highly expressed gene, *MITF*, was located within the amplified region (t -ratio = 17.36, P -value = 2.97×10^{-12} , adjusted threshold for significance = 3.96×10^{-6} ; Figs 1c, d). In a similar analysis restricted to just the eight NCI60 melanoma cell lines, the mean *MITF* expression ratio between the 3p-amplified (six samples) and non-amplified (two samples) classes exceeded that of nearly all randomly permuted class distinctions (data not shown). *MITF* encodes a basic helix–loop–helix/leucine zipper transcription factor required for development of the melanocyte lineage¹³, but not previously known to be the target of an acquired somatic mutation. These data raised the possibility that *MITF* might also function as a lineage-specific oncogene.

We then investigated *MITF* gene dosage in human tumours by performing quantitative polymerase chain reaction (PCR) on DNA derived from a series of melanocytic nevi, primary cutaneous melanomas, and melanoma metastases. *MITF* amplification was observed in 3 of 30 (10%) primary cutaneous and 7 of 32 (21%) metastatic tumours, but not in the ten benign nevi tested (Fig. 2a). *MITF* was the most consistently amplified gene in the 'core' amplified region defined by MALME-3M cells (Fig. 1d) and, when amplified to high levels, appeared to represent the epicentre of the amplicon (see,

¹Departments of Medical Oncology, Pediatric Oncology, Biostatistical Sciences and Melanoma Program in Medical Oncology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115, USA. ²Departments of Pathology, Brigham and Women's Hospital and Harvard Medical School, 75 Francis Street, Boston, MA 02115, USA. ³Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, MA 02115, USA. ⁴Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA. ⁵The Broad Institute of Harvard and MIT, 320 Charles Street, Cambridge, Massachusetts 02141, USA. ⁶Department of Pathology, Yale University School of Medicine, 310 Cedar Street, New Haven, Connecticut 06510, USA. ⁷Massachusetts General Hospital Center for Cancer Research and Department of Medicine, 149 13th Street, Charlestown, Massachusetts 02129, USA. ⁸DAID, Department of Dermatology, Medical University of Vienna, and Center of Molecular Medicine, Austrian Academy of Sciences, Währinger Gürtel 18–20, A-1090 Vienna, Austria.

for example, samples IM3 and MM5; Fig. 2c). Together, these data delineated *MITF* as the probable oncogene targeted by this genetic alteration.

Next, we examined nearly 200 tissue specimens derived from primary and metastatic melanomas by fluorescence *in situ* hybridization (FISH) using bacterial artificial chromosome (BAC) probes spanning the *MITF* locus (to detect extra-chromosomal amplification events). *MITF* amplification (range = 4–13 copies per cell; Fig. 3b and Supplementary Table 3) was detected in 2 of 19 (10.5%) cutaneous tumours and 27 of 160 (15.2%) metastatic samples. Once again, no amplification was detected in nine melanocytic nevi analysed. In metastatic melanoma, *MITF* amplification was associated with decreased 5-year survival (Fig. 3c; Kaplan–Meier log rank P value = 0.024), but not with other clinicopathologic parameters (Supplementary Figs 2 and 3). *MITF* protein levels were also analysed in the melanoma tissue microarray using automated quantitative analysis (AQUA) technology¹⁴. *MITF* amplification

correlated with a significantly increased mean *MITF* protein expression in metastatic disease (P = 0.019; Figs 3d–f). This association remained significant when expression was measured as a continuous variable in relation to *MITF* copy number (data not shown). Together, these observations implicate *MITF* amplification in the progression and lethality of a subset of human melanomas.

In melanocytes, activated MAP kinase triggers *MITF* phosphorylation at serine-73 (ref. 15), recruiting the transcriptional coactivator p300 (ref. 16) while simultaneously targeting *MITF* for ubiquitin-dependent proteolysis¹⁷. Normal melanocyte growth/differentiation may also require the coordinated actions of *MITF* and cyclin-dependent kinase (CDK) inhibitors such as p16 or p21 (refs 18, 19). On the other hand, aberrant MAP kinase signalling through activating *BRAF* or *NRAS* mutations may underlie a substantial percentage of melanomas^{20,21}. Interestingly, all NCI60 cell lines harbouring *MITF* copy gain also contained both the *BRAF*(V600E) mutation and p16 pathway inactivation²² (Supplementary Table 2). Thus,

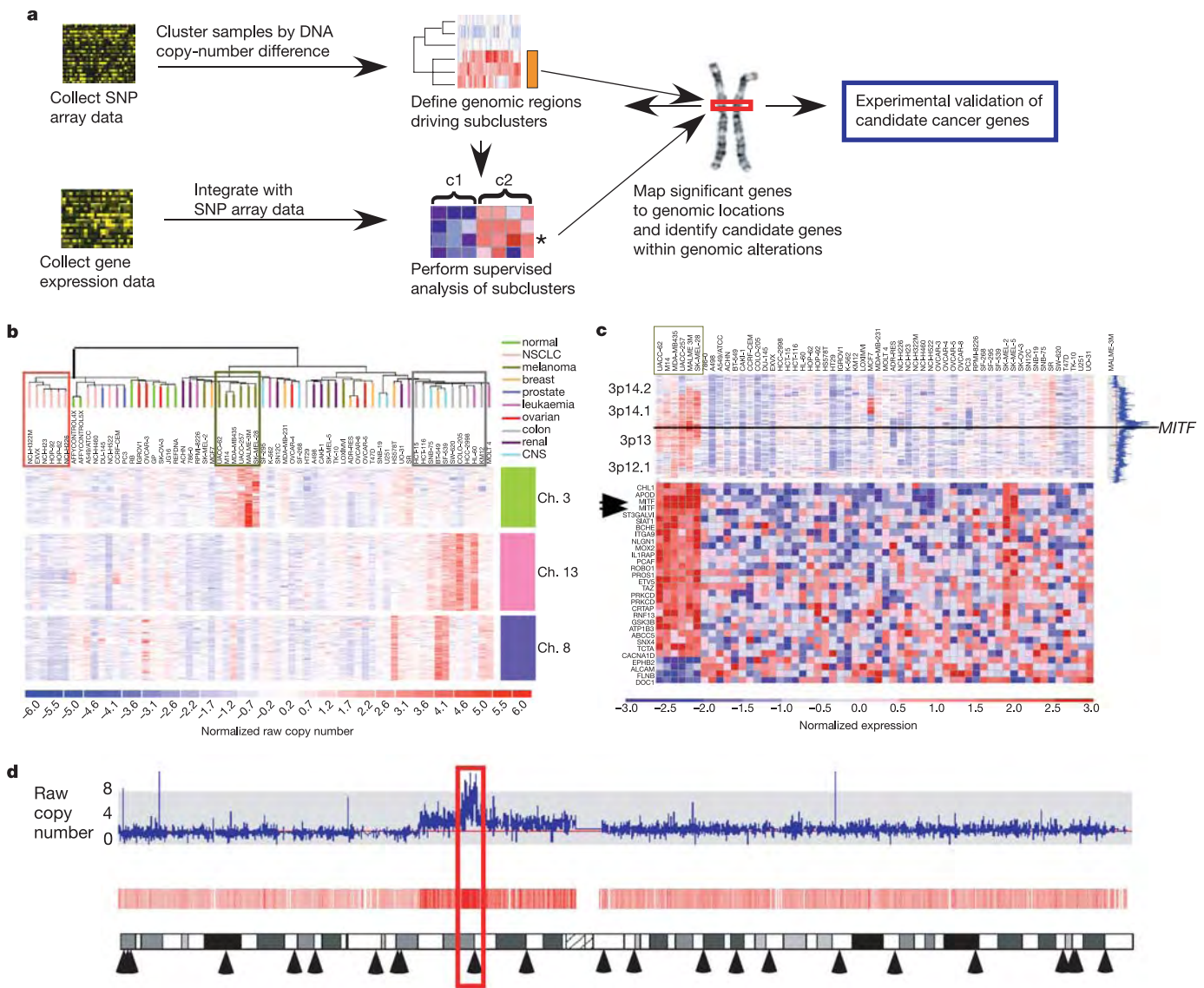


Figure 1 | Increased *MITF* expression associated with chromosome 3p amplification in melanoma cell lines. **a**, Schematic of the integrated genomic approach (see text). **b**, Hierarchical clustering of raw copy-number data from NCI60 cell lines and normal diploid controls shows cell line subclusters (coloured rectangles, top) and the corresponding chromosome specific SNP clusters (right). CNS, central nervous system tumour; NSCLC, non-small-cell lung cancer. **c**, Integration of copy-number and

gene-expression data. Shown are colourgrams of SNP copy number at 3p14–3p12 (top) and the significant gene-expression correlates from chromosome 3 following supervised analysis (bottom). **d**, Chromosome 3 copy-number values (top), SNP signal intensity colourgram (middle), cytoband map (bottom), significant genes (arrows), and the MALME-3M core amplicon (red box) are shown.

though BCL2 downregulation²³. Thus, these data suggest that deregulation of MITF through amplification or other mechanisms may preserve a critical lineage survival function in melanoma.

Melanocyte survival mechanisms that are modified by MITF may also contribute to melanoma chemoresistance^{23,25}. To investigate whether *MITF* copy gain correlated with drug resistance, we performed a supervised analysis of the available NCI60 pharmacological data, where chemical sensitivity to a large number of small molecules and natural products is known (see Supplementary Methods). 3p copy gain was associated with a significantly increased mean GI50 in 270 compounds (median 35 ± 39.7 compounds expected at random). This degree of chemoresistance was statistically significant within the overall NCI60 data set ($P = 0.016$; Supplementary Fig. 5a, b). An analysis restricted to just the eight NCI60 melanoma cell lines revealed a similar trend, though this did not reach statistical significance with the limited available permutations (Supplementary Fig. 5c). To examine the specific effect of MITF on melanoma drug resistance, MALME-3M melanoma cells (which contain the *MITF* amplicon) were infected with Ad-dnMITF at sublethal multiplicity of infection (MOI) and cultured in the presence of commonly used chemotherapeutic agents. MALME 3M cells infected with empty adenovirus showed pharmacologic profiles similar to wild-type controls (Supplementary Fig. 5d). However, cells expressing dnMITF were significantly more susceptible to inhibition by 20 μ M of both cisplatin and docetaxel after 72 h (Fig. 4h). Drug titration curves also showed a four- to fivefold decrease in the cellular growth inhibitory concentration (GI₅₀) of each agent in the presence of dnMITF (data not shown). Thus, reduction of MITF activity may sensitize melanomas to conventional chemotherapeutics.

Somatic alteration through gene amplification suggests that *MITF*

may be a member of a newly recognized 'lineage survival' or 'lineage addiction' oncogene subclass. Unlike oncogene addiction, where inappropriate gene-activation events render cells hyperdependent on particular pathways, dependency in this case might be pre-established during development and maintained in tumours through genetic aberrations. The androgen receptor (AR) is a prototype of this oncogene class that now includes *MITF*: both AR and MITF are required for the development and survival of their respective (prostate and melanocyte) lineages, maintained in cancers of these lineages, and amplified (or activated by mutation) in association with advanced disease. AR gene amplification typically occurs after androgen withdrawal therapy in prostate cancer patients. Although the cause of MITF amplification is not known, we speculate that increased MITF gene dosage (or deregulation by other means) may preserve its essential survival function in settings where the normal cues preserving MITF activity are lost. BRAF mutation and metastasis may constitute two such settings.

Although AR induces differentiation and growth arrest in non-transformed prostate epithelia in the presence of androgen, AR activation in the setting of p53 and Rb inactivation causes tumours in mice²⁶. Similarly, MITF induces both differentiation and growth arrest in non-transformed cells^{18,19}. Our results suggest that inactivation of p53 and Rb (the latter through loss of p16 activity) likewise enables MITF to transform human melanocytes in cooperation with mutated BRAF.

AR inhibition remains the most effective therapeutic intervention in prostate cancer, and emerging data suggests a continued dependency on AR in hormone-refractory disease²⁷. Extending the lineage survival oncogene analogy, patient selection based on MITF, BRAF and p16 pathway mutation status might identify responders or

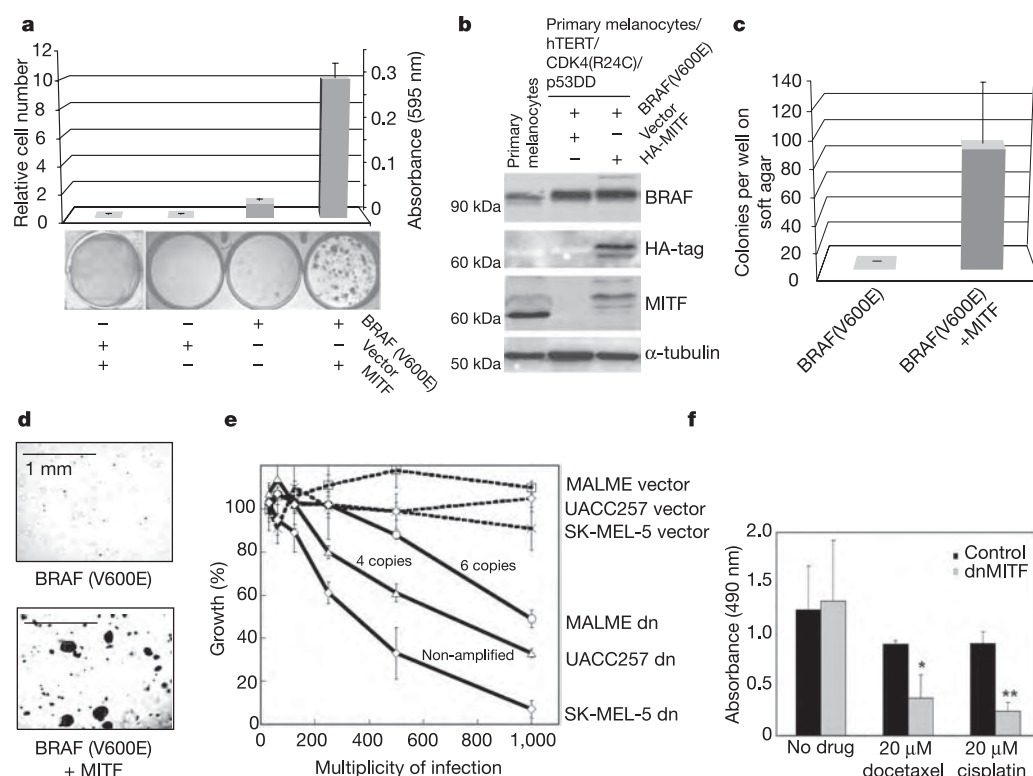


Figure 4 | A role for deregulated MITF in melanoma tumorigenesis and survival. **a**, MITF and BRAF(V600E) co-expression confers factor-independent growth in hTERT/CDK4(R24C)/p53DD melanocytes. Absorbance (mean and standard error) following crystal violet staining, and photographs at 3 weeks are shown. **b**, Extracts from hTERT/CDK4(R24C)/p53DD melanocytes expressing BRAF(V600E) ± HA-MITF were immunoblotted using antibodies against BRAF, MITF, HA-tag or α-tubulin. p53DD, dominant-negative p53. **c**, **d**, Growth of cells expressing

BRAF(V600E) ± HA-MITF in soft agar. Colonies (mean and standard error) and photographs taken at 8 weeks are shown (magnification $40\times$). **e**, Growth of melanoma cell lines expressing empty vector or dominant-negative MITF (dn) at 48 hours relative to uninfected controls. Means and standard deviations are shown; *MITF* copy number is indicated. **f**, Cell growth (mean and standard deviation, 72 h) of MALME-3M cells expressing dnMITF compared to uninfected controls, in the presence or absence of docetaxel or cisplatin (* $P < 0.05$, ** $P < 0.01$).

non-responders to the BRAF or CDK inhibitors currently in clinical development. Future combined genomic and functional studies may clarify tumour-survival mechanisms across many tumour types, thereby increasing biological understanding and providing new therapeutic possibilities.

METHODS

SNP array hybridization. Genomic DNA from 58 NCI60 cell lines was obtained from R. Camalier and D. Scudiero of the National Cancer Institute Developmental Therapeutics programme (NCI DTP). All DNAs were quantified by Picogreen (Molecular Probes) before SNP array analysis. SNP array data collection, normalization and copy-number determination are described in the Supplementary Information.

Hierarchical clustering using raw copy number data. Raw copy number data from CentXba arrays was filtered to reduce invariant SNPs using dChipSNP software (available at <http://www.dchip.org>) by adjusting parameters for the values of standard deviation/mean. Filtered SNPs were subjected to hierarchical clustering using the Pearson coefficient and average linkage method¹¹. Two-dimensional dendrograms (one for cell lines and one for SNPs) were generated (the full two-dimensional dendrogram is shown in the Supplementary Fig. S1). Nearly identical dendrograms were obtained for both CentHind and CentXba SNP data (not shown).

Supervised analysis of SNP array and gene expression data. The NCI60 cell lines were grouped into two classes based on the presence or absence of amplification at chromosome 3p14-3p13. We then performed supervised analyses using in-triplicate NCI60 gene expression data generated by Novartis on the Affymetrix U95v2 array platform (<http://dtp.nci.nih.gov/mtargets/madownload.html>). An extended description of this approach is presented in the Supplementary Information. To identify genes whose means differed significantly between classes, we performed 500 permutations of class labels using the *t*-test metric. *t* scores and corresponding *P* values were also calculated by standard methods, and *P*-values were adjusted to account for multiple hypotheses, as described in the Supplementary Information.

Quantitative, real-time PCR. Quantitative PCR was performed using a SYBR Green kit (Applied Biosystems) and either a PRISM 7500 sequence detector or the PRISM 7300 384-well sequence detector (Applied Biosystems) as described in the Supplementary Information.

FISH. Tissue samples used for interphase FISH analysis were derived from a melanoma tissue microarray described elsewhere²⁸. BAC clones RP11-584A6 and RP11-444P10 spanning the *MITF* locus were obtained from the Children's Hospital of Oakland Research Institute (CHORI). BAC DNA was prepared with the NucleoBond Plasmid Mini Kit (Clontech). Probe labelling with Digoxin-dUTP (Roche) was performed as described²⁹. Centromere probes were obtained from Vysis. The integrity and purity of all probes were verified by hybridization to metaphase spreads before tissue analysis. Tissue hybridization, washing, and bicolour detection were performed as described previously²⁹. *MITF* amplification was defined as the presence of four or more gene copies compared to two copies of the chromosome-3 centromere probe in at least 50 nuclei.

Kaplan–Meier analysis. Survival curves for 160 metastatic melanoma patients (analysed by FISH as described above) were calculated using Kaplan–Meier analysis with assessment of statistical significance using the Mantel–Cox log-rank test. The Cox proportional-hazards model was used for multivariate analysis to determine relative risk and independent significance. Analyses were performed with Statview 5.0.1 (SAS Institute). Patients were deemed uncensored if they died of melanoma within 5 years of their initial date of diagnosis. Survival was measured from the time of initial diagnosis. Clinicopathologic correlations were determined as described in the Supplementary Information.

AQUA of MITF protein expression. AQUA image acquisition and analysis¹⁴ was performed as described in the Supplementary Information.

Adenoviral vector construction. The dominant-negative *MITF* construct was generated as described in the Supplementary Information.

Cell line culture and adenoviral infections. NCI60 melanoma cell lines (MALME-3M, SK-MEL-5 and UACC257) were obtained from R. Camalier and D. Scudiero (NCI DTP) and were cultured in RPMI containing L-glutamine, 10% fetal bovine serum (FBS; HyClone), and antibiotic/antimycotic (Sigma). For adenoviral infections, 5,000 cells per well were plated in a 96-well plate. After adhering for 6–8 h, adenoviruses described in the text were added at the indicated MOI in serum-free RPMI containing 10 mM MgCl₂ for 30 min at 24 °C. After infection, cells were supplemented with RPMI containing FBS and cultured at 37 °C/5% CO₂ for up to 72 h before subsequent analyses (see below).

Preparation and retroviral transduction of primary melanocytes. Isolation, immortalization, and retroviral transduction of primary human melanocytes is described in the Supplementary Information.

Soft agar assays. Cells (10,000 per well) were seeded in 0.5% low-melting-point agarose in DMEM with 10% FBS, layered onto 0.8% agarose in DMEM/10% FBS. Colonies were stained with crystal violet, enumerated and photographed at 4 × magnification (40 × microscopy) eight weeks after seeding.

Pharmacologic data analysis. Pharmacologic data (–log₁₀[GI₅₀]) for 42,796 compounds was downloaded from the NCI website and analysed by supervised learning methods as described in the Supplementary Information. GI₅₀ is defined as the drug concentration achieving 50% growth inhibition relative to untreated cells.

Drug sensitivity and cell viability assays. MALME-3M cells were plated at 5,000 cells per well in a 96-well plate and infected at 500 MOI with adenovirus containing dominant-negative *MITF* (as described above). After 12 h, infected cells and controls were incubated with RPMI containing either 20 μM docetaxel or 20 μM cisplatin for 72 h. Cell viability following drug treatment was measured using the MTS-based CellTiter 96-cell proliferation assay (Promega).

Received 1 December 2004; accepted 19 April 2005.

1. Lindblad-Toh, K. *et al.* Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nature Biotechnol.* **18**, 1001–1005 (2000).
2. Lieberfarb, M. E. *et al.* Genome-wide loss of heterozygosity analysis from laser capture microdissected prostate cancer using single nucleotide polymorphic allele (SNP) arrays and a novel bioinformatics platform dChipSNP. *Cancer Res.* **63**, 4781–4785 (2003).
3. Mei, R. *et al.* Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res.* **10**, 1126–1137 (2000).
4. Zhao, X. *et al.* An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.* **64**, 3060–3071 (2004).
5. Bignell, G. R. *et al.* High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.* **14**, 287–295 (2004).
6. Stinson, S. F. *et al.* Morphological and immunocytochemical characteristics of human tumour cell lines for use in a disease-oriented anticancer drug screen. *Anticancer Res.* **12**, 1035–1053 (1992).
7. Ross, D. T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.* **24**, 227–235 (2000).
8. Roschke, A. V. *et al.* Karyotypic complexity of the NCI-60 drug-screening panel. *Cancer Res.* **63**, 8634–8647 (2003).
9. Nishizuka, S. *et al.* Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. *Proc. Natl Acad. Sci. USA* **100**, 14229–14234 (2003).
10. Monks, A., Scudiero, D. A., Johnson, G. S., Paull, K. D. & Sausville, E. A. The NCI anti-cancer drug screen: a smart screen to identify effectors of novel targets. *Anticancer Drug Des.* **12**, 533–541 (1997).
11. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
12. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
13. Steingrimsson, E., Copeland, N. G. & Jenkins, N. A. Melanocytes and the Microphthalmia Transcription Factor Network. *Annu. Rev. Genet.* **38**, 365–411 (2004).
14. Camp, R. L., Dolled-Filhart, M., King, B. L. & Rimm, D. L. Quantitative analysis of breast cancer tissue microarrays shows that both high and normal levels of HER2 expression are associated with poor outcome. *Cancer Res.* **63**, 1445–1448 (2003).
15. Hemesath, T. J., Price, E. R., Takemoto, C., Badalian, T. & Fisher, D. E. MAP kinase links the transcription factor Microphthalmia to c-Kit signalling in melanocytes. *Nature* **391**, 298–301 (1998).
16. Price, E. R. *et al.* Lineage-specific signalling in melanocytes. C-kit stimulation recruits p300/CBP to microphthalmia. *J. Biol. Chem.* **273**, 17983–17986 (1998).
17. Wu, M. *et al.* c-Kit triggers dual phosphorylations, which couple activation and degradation of the essential melanocyte factor Mi. *Genes Dev.* **14**, 301–312 (2000).
18. Loercher, A. E., Tank, E. M., Delston, R. B. & Harbour, J. W. *MITF* links differentiation with cell cycle arrest in melanocytes by transcriptional activation of *INK4A*. *J. Cell Biol.* **168**, 35–40 (2005).
19. Carreira, S. *et al.* *Mitf* cooperates with *Rb1* and activates *p21Cip1* expression to regulate cell cycle progression. *Nature* **433**, 764–769 (2005).
20. Davies, H. *et al.* Mutations of the *BRAF* gene in human cancer. *Nature* **417**, 949–954 (2002).
21. Jafari, M. *et al.* Analysis of ras mutations in human melanocytic lesions: activation of the ras gene seems to be associated with the nodular type of human malignant melanoma. *J. Cancer Res. Clin. Oncol.* **121**, 23–30 (1995).
22. Kubo, A. *et al.* The p16 status of tumour cell lines identifies small molecule inhibitors specific for cyclin-dependent kinase 4. *Clin. Cancer Res.* **5**, 4279–4286 (1999).
23. McGill, G. G. *et al.* Bcl2 regulation by the melanocyte master regulator *Mitf* modulates lineage survival and melanoma cell viability. *Cell* **109**, 707–718 (2002).

24. Du, J. *et al.* Critical role of CDK2 for melanoma growth linked to its melanocyte-specific transcriptional regulation by MITF. *Cancer Cell* **6**, 565–576 (2004).
25. Chu, W. *et al.* Tyrosinase-related protein 2 as a mediator of melanoma specific resistance to cis-diamminedichloroplatinum(II): therapeutic implications. *Oncogene* **19**, 395–402 (2000).
26. Berger, R. *et al.* Androgen-induced differentiation and tumorigenicity of human prostate epithelial cells. *Cancer Res.* **64**, 8867–8875 (2004).
27. Chen, C. D. *et al.* Molecular determinants of resistance to antiandrogen therapy. *Nature Med.* **10**, 33–39 (2004).
28. Berger, A. J. *et al.* Automated quantitative analysis (AQUA) of HDM2 expression in malignant melanoma shows association with early stage disease and improved outcome. *Cancer Res.* (in the press).
29. Rubin, M. A. *et al.* Overexpression, amplification, and androgen regulation of TPD52 in prostate cancer. *Cancer Res.* **64**, 3814–3822 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Scudiero and R. Camalier for provision of NCI60 cell lines and DNAs, O. Kabbarah and L. Chin for discussions and provision of reagents, F. Chen and C. Ladd-Acosta for excellent technical assistance, L. Ziaugra and S. Gabriel for assistance with the BRAF(V600E) genotyping assay, and M. Loda for expert advice. This work was supported by grants from the National Institutes of Health (L.A.G., M.A.R., D.L.R. and D.E.F.), the Swedish Wenner-Gren Foundation (H.R.W.), the Center of Molecular Medicine, Austrian Academy of Sciences (S.N.W.), the Howard Hughes Medical Institute (T.R.G.), the American Cancer Society (M.L.M.), the Flight Attendant Medical Research Institute (M.L.M.), the Doris Duke Foundation (D.E.F.), the Tisch Family Foundation (W.R.S.), and the Damon Runyon Cancer Research Foundation (W.R.S.).

Author Information The GEO accession number is GSE2520. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to W.R.S. (William_Sellers@dfci.harvard.edu).

Rac1b and reactive oxygen species mediate MMP-3-induced EMT and genomic instability

Derek C. Radisky¹, Dinah D. Levy¹, Laurie E. Littlepage², Hong Liu¹, Celeste M. Nelson¹, Jimmie E. Fata¹, Devin Leake³, Elizabeth L. Godden³, Donna G. Albertson⁴, M. Angela Nieto⁵, Zena Werb² & Mina J. Bissell¹

The tumour microenvironment can be a potent carcinogen, not only by facilitating cancer progression and activating dormant cancer cells, but also by stimulating tumour formation¹. We have previously investigated stromelysin-1/matrix metalloproteinase-3 (MMP-3), a stromal enzyme upregulated in many breast tumours², and found that MMP-3 can cause epithelial–mesenchymal transition (EMT) and malignant transformation in cultured cells^{3–5}, and genomically unstable mammary carcinomas in transgenic mice³. Here we explain the molecular pathways by which MMP-3 exerts these effects: exposure of mouse mammary epithelial cells to MMP-3 induces the expression of an alternatively spliced form

of Rac1, which causes an increase in cellular reactive oxygen species (ROS). The ROS stimulate the expression of the transcription factor Snail and EMT, and cause oxidative damage to DNA and genomic instability. These findings identify a previously undescribed pathway in which a component of the breast tumour microenvironment alters cellular structure in culture and tissue structure *in vivo*, leading to malignant transformation.

Cancer is characterized by a progressive series of alterations that disrupt cell and tissue homeostasis. Whereas many of these alterations can be induced by specific mutations, faulty signals from the microenvironment also can act as inducers of tumour development

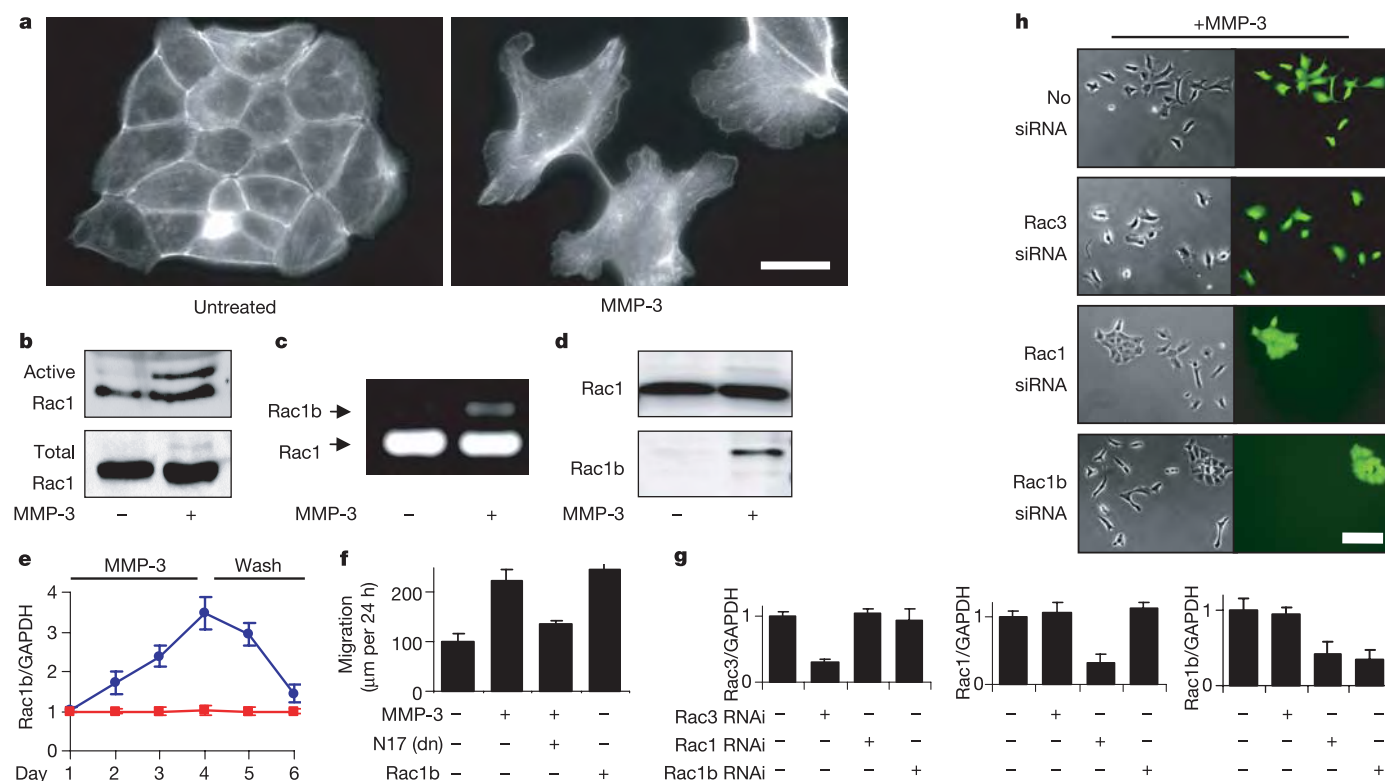


Figure 1 | MMP-3 induces EMT through Rac1b. **a**, MMP-3-induced alterations in actin cytoskeleton. Scale bar, 25 μ m. **b**, Analysis of active and total levels of Rac. **c**, RT-PCR of Rac1 and Rac1b. **d**, Rac1b protein expression. **e**, Rac1b transcript levels in response to MMP-3 treatment (days 1–4) and washout (days 5–6); blue circles, treated; red squares, untreated.

f, Cell motility assessed by scratch assay. dn, dominant-negative. **g**, Quantification of knockdown of endogenous gene expression. **h**, Selective knockdown of Rac1b inhibits MMP-3-induced cell scattering. Scale bar, 25 μ m. For all graphs, error bars represent s.e.m.

¹Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. ²Department of Anatomy, University of California, San Francisco, California 94143, USA. ³Dharmacon Inc., Lafayette, Colorado 80026, USA. ⁴Department of Laboratory Medicine, and Comprehensive Cancer Center, University of California, San Francisco, California 94143, USA. ⁵Department of Developmental Neurobiology, Instituto de Neurociencias de Alicante, 03550 San Juan de Alicante, Spain.

and progression¹. MMPs are prominent contributors to such micro-environmental signals, because these proteolytic enzymes degrade structural components of the extracellular matrix (ECM), permitting tumour invasion and metastasis. Additionally, MMPs can release cell-bound inactive precursor forms of growth factors, degrade cell-cell and cell-ECM adhesion molecules, activate precursor zymogen forms of other MMPs, and inactivate inhibitors of MMPs and other proteases⁶. Our observations that MMP-3 can induce transformation in mammary epithelial cells in culture^{4,5,7} and in transgenic mice³ prompted investigations into the underlying molecular mechanisms. Here we show that MMP-3 can lead directly to genomic instability and EMT, and identify the pathway that mediates these events.

Induction of EMT by treatment of SCp2 mouse mammary epithelial cells with MMP-3 is associated with a loss of intact E-cadherin, increased motility and invasiveness, downmodulation of epithelial markers, and upregulation of mesenchymal markers (refs 4, 5; Supplementary Fig. 1a,b), through a process that is initially reversible (refs 4, 5; Supplementary Fig. 1c). The MMP-3-induced alteration of the F-actin cytoskeleton (Fig. 1a) indicated the possible involvement of members of the Rho GTPase family, and although the activities of RhoA and Cdc42 were unchanged (not shown), we were intrigued by an additional band in the Rac activity assay of MMP-3-treated cells (Fig. 1b). A highly activated splice isoform of Rac1, designated Rac1b, containing 57 additional nucleotides that result in an in-frame insertion of 19 additional amino acid residues, was discovered recently in breast and colorectal tumours^{8,9} and has transforming characteristics when exogenously expressed in cultured cells¹⁰. We identified the additional Rac band induced by MMP-3 as Rac1b by polymerase chain reaction with reverse transcription (RT-PCR) (Fig. 1c) and through the use of an antibody raised against the mouse Rac1b insertion sequence (Fig. 1d); we also found that

induction of Rac1b by treatment with MMP-3 was initially reversible (Fig. 1e). We determined that the activity of Rac1b was required for the MMP-3-induced alterations in vimentin expression (Supplementary Fig. 2) and for MMP-3-induced motility (Fig. 1f), because dominant-negative Rac1N17 attenuated the effects of MMP-3, and expression of Rac1b could substitute for MMP-3 (Fig. 1f).

We also evaluated the relationship between the induction of Rac1b and downstream EMT by specific transcript knockdown with short interfering RNA (siRNA). SCp2 cells were co-transfected transiently with yellow fluorescent protein (YFP), YFP-Rac1 or YFP-Rac1b, and either no siRNA, siRNA targeting Rac3 (as negative control), siRNA targeting Rac1 (which also targets Rac1b) or siRNA selectively targeting the splice insertion sequence in Rac1b. We found that Rac1 siRNA blocked the expression of co-transfected YFP-Rac1b, and that the specific Rac1b siRNA blocked the expression of co-transfected YFP-Rac1b only, not YFP-Rac1 (Supplementary Fig. 3); none of the siRNAs affected the expression of co-transfected YFP. The effect on endogenous gene expression levels was consistent with effective knockdown of all transiently transfected cells (about 70% transfection efficiency) and showed the following: Rac1 siRNA inhibits the expression of both Rac1 and Rac1b but does not affect Rac3; Rac1b siRNA selectively targets Rac1b and does not affect Rac1 or Rac3; and Rac3 siRNA selectively targets Rac3 and does not affect expression of Rac1 or Rac1b (Fig. 1g). When SCp2 cells were transiently co-transfected with YFP and either no siRNA or siRNA targeting Rac3, Rac1/Rac1b or Rac1b, and then treated with MMP-3 for 4 days, we observed that siRNA for Rac1/Rac1b or Rac1b inhibited MMP-3-induced cell motility in the co-transfected colonies, whereas siRNA targeting Rac3 had no effect (Fig. 1h).

How can increased Rac activity lead to the diverse alterations induced by MMP-3? Previous studies^{11,12} showed that active Rac can

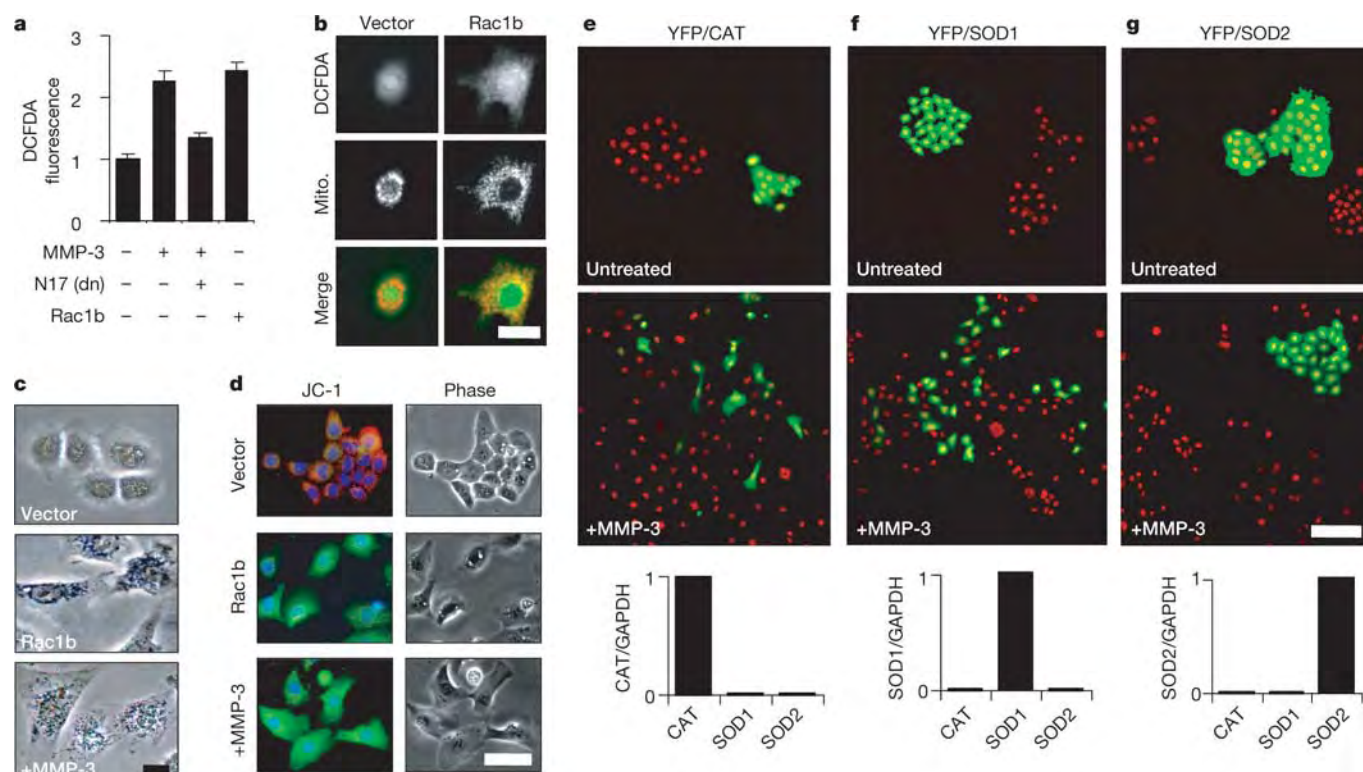


Figure 2 | MMP-3/Rac1b stimulate mitochondrial production of ROS.

a, Cellular ROS levels assessed by DCFDA. dn, dominant-negative. Error bars represent s.e.m. **b**, Mitochondrial pattern of DCFDA fluorescence. Scale bar, 25 μ m. **c**, Precipitation of nitroblue tetrazolium. Scale bar, 15 μ m. **d**, Mitochondrial depolarization assessed with JC-1. Scale bar, 50 μ m.

e-g, Cells co-transfected with EYFP and catalase (CAT; **e**), superoxide dismutase 1 (SOD1; **f**) or superoxide dismutase 2 (SOD2; **g**) and then cultured in the absence (top) or presence (middle) of MMP-3 for 6 days. Green, EYFP fluorescence; red, nuclei. Graphs at bottom show gene transcript levels in transfected cell populations. Scale bar, 100 μ m.

stimulate the production and release of mitochondrial superoxide into the cytoplasm. The production of excess superoxide can cause oxidative DNA damage and genomic instability¹³, transform cells in culture¹⁴ and potentiate tumour progression¹⁵; in addition, superoxide is readily converted to other forms of ROS that stimulate further tumorigenic processes^{15–17}. We found that treatment with MMP-3 or the expression of Rac1b produced increases in cellular ROS, as assessed by the fluorophore dichlorodihydrofluorescein diacetate (DCFDA), and that the expression of Rac1N17 attenuated the induction of ROS by MMP-3 (Fig. 2a). The DCFDA fluorescence partly colocalized with a mitochondrial marker protein (Fig. 2b), and the identity of the induced ROS as mitochondrial superoxide was indicated by the staining pattern of nitroblue tetrazolium (Fig. 2c), which forms an insoluble blue formazan in the presence of superoxide¹², and by the altered fluorescence pattern of cells stained with JC-1, in which the punctate red mitochondrial staining of the J-aggregate of JC-1 was replaced by diffuse cytoplasmic green staining of the monomeric form (Fig. 2d), consistent with the dissipation of membrane potential after the mitochondrial production of superoxide^{12,18}. To determine whether the induction of mitochondrial superoxide by MMP-3/Rac1b was essential for the induction of EMT, we co-transfected cells with expression plasmids encoding YFP and either catalase, superoxide dismutase 1 (SOD1) or SOD2. Catalase stimulates the decomposition of H₂O₂ into water and molecular oxygen, whereas SOD1 and SOD2 convert superoxide into H₂O₂ and molecular oxygen; catalase and SOD1 are cytoplasmic enzymes, whereas SOD2 is localized to the mitochondria. These experiments demonstrated that YFP/SOD2 cells were resistant to MMP-3-induced scattering (Fig. 2g), whereas YFP/catalase and YFP/SOD1 cells responded in a similar fashion to adjacent untransfected cells (Fig. 2e, f).

ROS can alter gene expression^{15–17} and stimulate cell invasiveness¹⁹, and we found that the ROS-quenching agent *N*-acetyl cysteine (NAC) effectively inhibited the MMP-3-induced downregulation of epithelial cytokeratins (Fig. 3a) and the upregulation of mesenchymal vimentin (Fig. 3g). NAC also inhibited MMP-3-induced cell

motility, invasion and morphological alterations (not shown). The induction of EMT involves the coordinated regulation of many genes²⁰; here we focused on MMP-3/ROS-dependent alterations in the expression levels of transcriptional regulatory proteins that mediate EMT. We determined that MMP-3 enhances expression of the transcription factor Snail^{21,22} (Fig. 3c), and that this effect could be blocked by treatment with NAC, or induced in the absence of MMP-3 by elevating ROS levels with H₂O₂ or by the expression of Rac1b (Fig. 3b). Expression of Snail in SCp2 cells was sufficient to induce EMT: induction caused the downmodulation of E-cadherin transcript and protein levels (Fig. 3d, e) and led to cell scattering comparable to that induced by MMP-3 or H₂O₂ (Fig. 3f). We also found that whereas MMP-3, Rac1b, H₂O₂ or Snail can stimulate the expression of mesenchymal vimentin (Fig. 3g), only MMP-3 could stimulate the expression of Rac1b (Fig. 3h). When combined with the data presented in Figs 1b–f and 2a–d, these results show that treatment with MMP-3 stimulates the expression of Rac1b, which causes increases in cellular ROS, leading in turn to an increased expression of Snail and to EMT.

We had previously found that tumours in the MMP-3-expressing transgenic mice showed common patterns of genomic rearrangements³, indicating that MMP-3 might lead to genomic instability in target epithelial cells *in vivo*. Given the known genotoxic effects of ROS, we investigated the effects of MMP-3-induced ROS on the integrity of the genome under defined conditions in culture. To test for DNA damage, we used fluorescein isothiocyanate (FITC)-conjugated avidin, because this reagent binds to 8-oxodeoxyguanosine, an oxidative DNA lesion with a structural similarity to biotin²³. Cells treated with MMP-3 showed significantly increased FITC-avidin nuclear staining (Fig. 4a) that was blocked by preincubation with an oligonucleotide containing 8-oxodeoxyguanosine (but not with a control oligonucleotide; not shown), by inhibiting the proteolytic activity of MMP-3 with GM6001, or by treatment with NAC (Fig. 4b). To test for induction of genomic instability, we assayed for increased resistance of MMP-3-treated SCp2 mouse mammary epithelial cells to *N*-(phosphonacetyl)-L-aspartate (PALA)²⁴, because resistance to

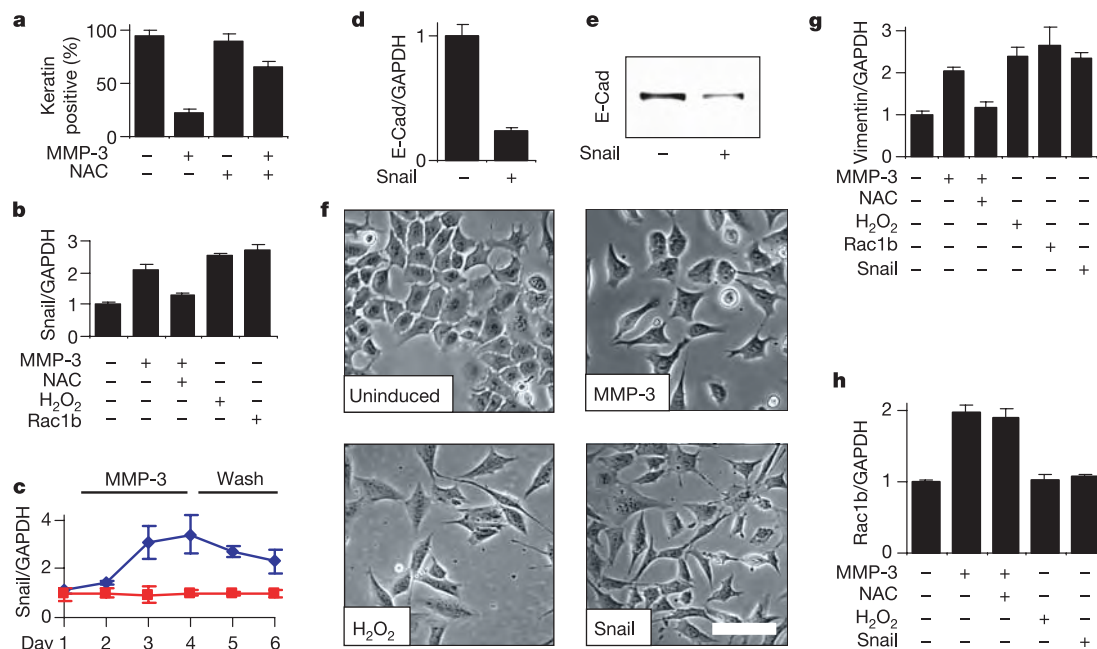


Figure 3 | MMP-3-induced EMT is dependent on ROS. **a**, NAC inhibits MMP-3-induced downregulation of epithelial cytokeratin protein levels. **b**, Induction of Snail by MMP-3, and dependence on ROS. **c**, Snail transcript levels in response to MMP-3 treatment (days 1–4) and washout (days 5–6); blue diamonds, treated; red squares, untreated. **d**, **e**, Exogenous expression

of Snail in SCp2 cells reduces E-cadherin transcript (**d**) and protein levels (**e**). **f**, Cell scattering induced by treatment with MMP-3 or H₂O₂, or by exogenous expression of Snail. Scale bar, 50 μ m. **g**, **h**, ROS and Snail dependence of vimentin (**g**) and Rac1b (**h**) expression. For all graphs, error bars represent s.e.m.

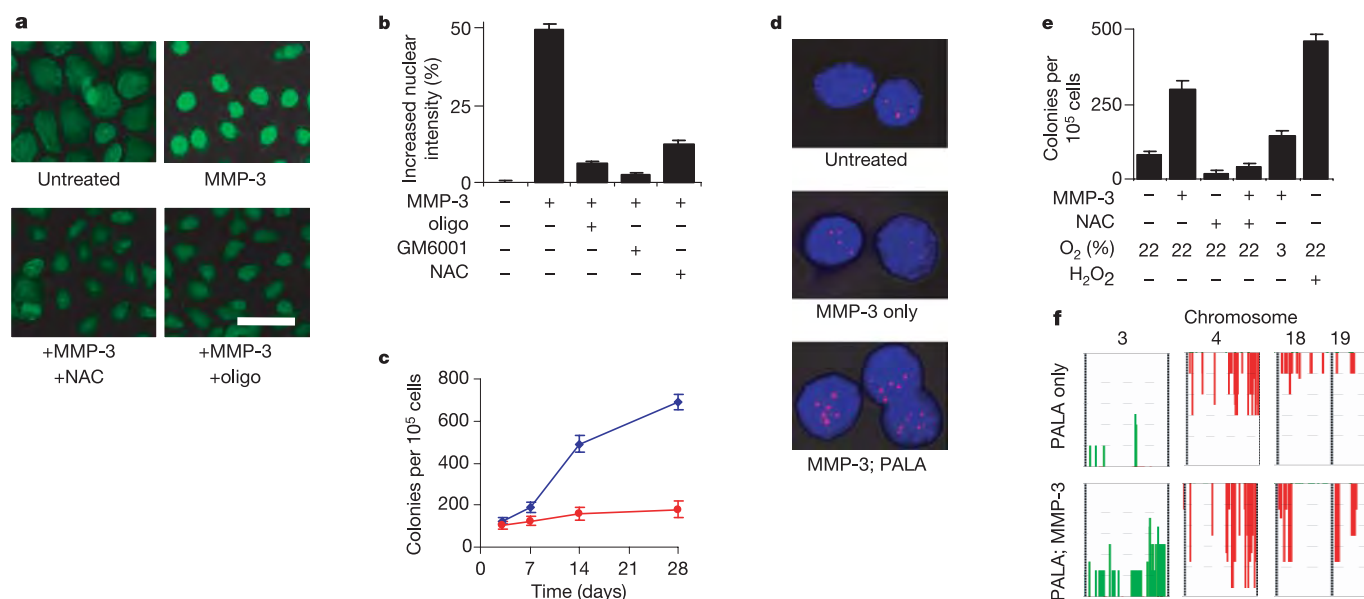


Figure 4 | MMP-3-induced ROS cause DNA damage and genomic instability. **a**, **b**, 8-Oxoguanosine-induced treatment with MMP-3 (**a**; scale bar, 50 μ m) and quantification of increased nuclear staining relative to untreated (**b**; error bars, 95% confidence intervals). **c**, Induction of PALA resistance by MMP-3 (blue diamonds, MMP-3; red squares, untreated). **d**, Fluorescence *in situ* hybridization of the CAD gene locus (red). **e**, ROS

and oxygen dependence of PALA resistance induced by 14 days of treatment with MMP-3. **f**, Frequency plots of comparative genomic hybridization analyses of cells grown in the absence (top) or presence (bottom) of MMP-3, and then selected with PALA. Green, amplifications; red, deletions. For all graphs, error bars represent s.e.m.

PALA is acquired through amplification of the *CAD* gene, which encodes carbamoylphosphate synthetase/aspartate carbamyltransferase/dihydroorotase²⁵. Exposure to MMP-3 led to a progressive increase in the fraction of cells that had acquired PALA resistance (Fig. 4c), an increase that was due to amplification of the *CAD* locus (Fig. 4d). This effect could also be inhibited by treatment with NAC or by culturing under reduced oxygen tension, and could be reproduced in the absence of MMP-3 by treatment with H₂O₂ (Fig. 4e). That the genomic instability induced by MMP-3 was not limited to the *CAD* locus was shown by comparative genomic hybridization analysis, because many additional genomic amplifications and deletions were found in MMP-3-treated cells (Fig. 4f), including characteristic alterations previously observed in tumours derived from the MMP-3 transgenic mice³.

Our results show that a key event in the MMP-3-induced malignant transformation of SCp2 cells is the induction of Rac1b, an alternative splice isoform of Rac1 that was initially identified in breast and colon cancers^{8,9}. Many oncogenic splice isoforms are induced in cancers²⁶, and although most of these produce proteins that lack key functional domains and act by sequestering factors involved in tumour suppressive pathways into nonfunctional complexes, Rac1b is unusual in that it becomes more highly activated^{27,28}. The fact that Rac1b is the only apparent splice isoform of Rac1 found in MMP-3-treated cells is also notable, because Rac1b is also the only apparent splice isoform in breast cancer cells⁹.

How does treatment with MMP-3 lead to alternative splicing of Rac1b? It is clear that the extracellular proteolytic activity of MMP-3 is essential (Supplementary Fig. 4). We have shown that MMP-3 effectively cleaves E-cadherin, resulting in a loss of cell–cell adhesions and the relocalization of transcriptionally active β -catenin to the nucleus (refs 4, 5; not shown). It is important to note that MMP-3 is not the only protease capable of initiating this pathway, as we have found that MMP-9 (but not MMP-2) can substitute for MMP-3 in our experimental system (not shown), and MMP-7 and MMP-14 are also known to induce tumours when expressed in transgenic mice². Furthermore, MMPs are not the only microenvironmental components implicated in tumour induction or progression. Oncogenic

properties have also been attributed to transforming growth factor- β , growth factors, and hormones, and the tumour-promoting activities of chronic inflammation are well known¹. Our investigations of MMP-3 show how this factor can directly stimulate phenotypic and genotypic malignant transformation in normally functioning cells. We expect that similar or parallel pathways may be induced by other elements of the tumour microenvironment, and we suspect that such mechanisms may be much more relevant for the generation of genomic instability than predicted by current models of tumour progression.

METHODS

Cell culture, antibodies and plasmids. Cell culture was performed as described previously^{4,5}; for gene repression, a 5 mg ml⁻¹ stock solution of tetracycline in 100% ethanol was diluted 1:1,000 into culture medium and changed daily. To stimulate cells with MMP-3, we used medium that had been conditioned by SCp2 cells containing the tetracycline-regulated, autoactivated MMP-3 construct^{4,5} with expression induced by growth in the absence of tetracycline; conditioned medium from cells repressed by treatment with tetracycline was used for controls. This conditioned medium was analysed by zymography to verify that MMP-3 was the only MMP being expressed, and that EMT was induced by extracellular proteolytic activity (Supplementary Fig. 4). Except as otherwise indicated, cells were incubated for 4 days in the presence of conditioned medium containing MMP-3 and for 7 days with 25 μ M H₂O₂. NAC was used at a concentration of 10 mM. Antibodies against cytokeratin and vimentin were described previously^{4,5}. Rac antibody was from Upstate, and the Rac1b antibody was raised against the peptide Ac-CGKDRPSRGKDKPIA-amide (antibody validation in Supplementary Fig. 5). Human catalase complementary DNA was obtained from R. Arnold, and human SOD1 and SOD2 cDNA were obtained from T.-T. Huang. SOD1, SOD2 and catalase were cloned into pcDNA3.1 expression vectors; all other constructs were subcloned into the tetracycline-repressible expression system used previously for the expression of MMP-3 (described in refs 4, 5). Rac1 and Rac1b were cloned from SCp2 cDNA and expressed as unmodified proteins or as fused with YFP. Rac1V12 and Rac11N17 mutants of Rac1 (Supplementary Figs 6 and 7), and also the catalytically inactive E217A mutant of MMP-3, were generated with the Quickchange mutagenesis kit (Stratagene); the sequences of all modified plasmids were verified. Transcript levels were assessed with RT–PCR by isolating RNA (Tri-pure; Roche Diagnostics), synthesizing cDNA and performing quantitative, real-time PCR (Lightcycler,

Roche Diagnostics); all these experiments were normalized to GAPDH. For analysis of Rac1 and Rac1b (Fig. 1c), oligonucleotide primers that hybridize to sequences flanking the splice insertion site were used; for specific analysis of Rac1b (Figs 1d and 3h), oligonucleotide primers specific for the Rac1b splice isoform were used.

Rho GTPase assays. Cells were lysed in glutathione *S*-transferase (GST)-Fish buffer (10% glycerol, 50 mM Tris-HCl pH 7.4, 100 mM NaCl, 1% Nonidet P40, 2 mM MgCl₂, 10 µg ml⁻¹ leupeptin, 10 µg ml⁻¹ pepstatin, 10 µg ml⁻¹ aprotinin, 10 µg ml⁻¹ E64 and 1 mM Pefabloc). Equal amounts of protein supernatants were incubated on ice for 45 min with GST-PAK-CD (Rac-binding domain and Cdc42-binding domain) or GST-C21 (Rho-binding domain) fusion protein-coated Sepharose beads. The beads were washed, eluted in sample buffer, and then analysed by SDS-polyacrylamide-gel electrophoresis and western blotting with antibodies against Rac, Cdc42 and Rho. Dominant-negative and constitutively active Rac1 expression constructs were provided by D. Kalman. Rac1 and Rac3 siRNA were SMARTpool reagents (Dharmacon), whereas Rac1b siRNA used the sequence 5'-UGGAGACACAUGUGGUAAGAUAGA-3'; siRNAs were transfected into SCp2 cells with Lipofectamine 2000 (Gibco) in accordance with the manufacturer's protocols. For analysis of endogenous gene knockdown, RNA was harvested after 24 h and analysed by RT-PCR with primer pairs selective for Rac1, Rac1b or Rac3. For MMP-3-induced EMT, siRNA mixtures were co-transfected with YFP-C1 and then treated with MMP-3 for 4 days; they were then evaluated for scatter of fluorescent (co-transfected with YFP and siRNA) and nonfluorescent (non-transfected control) colonies.

8-Oxodeoxyguanosine analyses and genomic instability assays. To measure ROS concentrations, cells were incubated in the dark with 50 mM DCFDA (Molecular Probes) for 30 min in serum-free and phenol-red-free medium. For 8-oxodeoxyguanosine analysis, a modification of published techniques^{23,29} was used: cells were fixed in methanol (20 min, -20 °C), permeabilized with TBST (Tris-buffered saline containing 0.1% Triton X-100; 15 min, 25 °C), blocked for non-specific binding (TBST containing 15% fetal calf serum; 2 h, 25 °C), and stained with 15 µg ml⁻¹ FITC-conjugated avidin (Sigma; 1 h, 37 °C). To verify the specificity of staining, FITC-avidin was preincubated with a tenfold excess of either the blocking oligonucleotide 5'-GAA CTA GTG ATC CCC CGG GTC GC-3' (where °G is 8-oxodeoxyguanosine) or the control oligonucleotide 5'-GAA CTA GTG ATC CCC CGG GTC GC-3'. Images were captured with a Nikon Diaphot 300 microscope and Spot RT camera and software (Technical Instruments). Fluorescence was quantified with IMAGEJ (<http://rsb.info.nih.gov/ij/index.html>). For DCFDA staining, cellular fluorescence was quantified; for FITC-avidin staining, nuclear fluorescence was measured (with a 4,6-diamidino-2-phenylindole image mask). More than 250 measurements were made for each data point. JC-1 and nitroblue tetrazolium labelling were performed essentially as in ref. 12. The PALA assay and the comparative genomic hybridization analyses were performed using modifications of previously published protocols^{24,25,30}, as described in the Supplementary Information.

Received 22 February; accepted 25 April 2005.

- Bissell, M. J. & Radisky, D. Putting tumours in context. *Nature Rev. Cancer* **1**, 46–54 (2001).
- Sternlicht, M. D. & Werb, Z. How matrix metalloproteinases regulate cell behaviour. *Annu. Rev. Cell Dev. Biol.* **17**, 463–516 (2001).
- Sternlicht, M. D. et al. The stromal proteinase MMP3/stromelysin-1 promotes mammary carcinogenesis. *Cell* **98**, 137–146 (1999).
- Lochter, A. et al. Misregulation of stromelysin-1 expression in mouse mammary tumour cells accompanies acquisition of stromelysin-1-dependent invasive properties. *J. Biol. Chem.* **272**, 5007–5015 (1997).
- Lochter, A. et al. Matrix metalloproteinase stromelysin-1 triggers a cascade of molecular alterations that leads to stable epithelial-to-mesenchymal conversion and a premalignant phenotype in mammary epithelial cells. *J. Cell Biol.* **139**, 1861–1872 (1997).
- Egeblad, M. & Werb, Z. New functions for the matrix metalloproteinases in cancer progression. *Nature Rev. Cancer* **2**, 161–174 (2002).
- Boudreau, N., Simpson, C. J., Werb, Z. & Bissell, M. J. Suppression of ICE and apoptosis in mammary epithelial cells by extracellular matrix. *Science* **267**, 891–893 (1995).
- Jordan, P., Brazao, R., Boavida, M. G., Gespach, C. & Chastre, E. Cloning of a novel human Rac1b splice variant with increased expression in colorectal tumors. *Oncogene* **18**, 6835–6839 (1999).
- Schnelzer, A. et al. Rac1 in human breast cancer: overexpression, mutation analysis, and characterization of a new isoform, Rac1b. *Oncogene* **19**, 3013–3020 (2000).
- Singh, A. et al. Rac1b, a tumour associated, constitutively active Rac1 splice variant, promotes cellular transformation. *Oncogene* **23**, 9369–9380 (2004).
- Kheradmand, F., Werner, E., Tremble, P., Symons, M. & Werb, Z. Role of Rac1 and oxygen radicals in collagenase-1 expression induced by cell shape change. *Science* **280**, 898–902 (1998).
- Werner, E. & Werb, Z. Integrins engage mitochondrial function for signal transduction by a mechanism dependent on Rho GTPases. *J. Cell Biol.* **158**, 357–368 (2002).
- Samper, E., Nicholls, D. G. & Melov, S. Mitochondrial oxidative stress causes chromosomal instability of mouse embryonic fibroblasts. *Aging Cell* **2**, 277–285 (2003).
- Suh, Y. A. et al. Cell transformation by the superoxide-generating oxidase Mox1. *Nature* **401**, 79–82 (1999).
- Droge, W. Free radicals in the physiological control of cell function. *Physiol. Rev.* **82**, 47–95 (2002).
- Puri, P. L. et al. A myogenic differentiation checkpoint activated by genotoxic stress. *Nature Genet.* **32**, 585–593 (2002).
- Finkel, T. Oxidant signals and oxidative stress. *Curr. Opin. Cell Biol.* **15**, 247–254 (2003).
- Madesh, M. & Hajnoczky, G. VDAC-dependent permeabilization of the outer mitochondrial membrane by superoxide induces rapid and massive cytochrome c release. *J. Cell Biol.* **155**, 1003–1015 (2001).
- Mori, K., Shibamura, M. & Nose, K. Invasive potential induced under long-term oxidative stress in mammary epithelial cells. *Cancer Res.* **64**, 7464–7472 (2004).
- Kalluri, R. & Neilson, E. G. Epithelial-mesenchymal transition and its implications for fibrosis. *J. Clin. Invest.* **112**, 1776–1784 (2003).
- Nieto, M. A. The snail superfamily of zinc-finger transcription factors. *Nature Rev. Mol. Cell Biol.* **3**, 155–166 (2002).
- Thiery, J. P. Epithelial-mesenchymal transitions in tumour progression. *Nature Rev. Cancer* **2**, 442–454 (2002).
- Struthers, L., Patel, R., Clark, J. & Thomas, S. Direct detection of 8-oxodeoxyguanosine and 8-oxoguanine by avidin and its analogues. *Anal. Biochem.* **255**, 20–31 (1998).
- Johnson, R. K., Inouye, T., Goldin, A. & Stark, G. R. Antitumor activity of N-(phosphonacetyl)-L-aspartic acid, a transition-state inhibitor of aspartate transcarbamylase. *Cancer Res.* **36**, 2720–2725 (1976).
- Wahl, G. M., Padgett, R. A. & Stark, G. R. Gene amplification causes overproduction of the first three enzymes of UMP synthesis in N-(phosphonacetyl)-L-aspartate-resistant hamster cells. *J. Biol. Chem.* **254**, 8679–8689 (1979).
- Shin, C. & Manley, J. L. Cell signalling and the control of pre-mRNA splicing. *Nature Rev. Mol. Cell Biol.* **5**, 727–738 (2004).
- Matos, P., Collard, J. G. & Jordan, P. Tumor-related alternatively spliced Rac1b is not regulated by Rho-GDP dissociation inhibitors and exhibits selective downstream signalling. *J. Biol. Chem.* **278**, 50442–50448 (2003).
- Fiorenza, D. et al. Alternative splicing of Rac1 generates Rac1b, a self-activating GTPase. *J. Biol. Chem.* **279**, 4743–4749 (2004).
- Neumann, C. A. et al. Essential role for the peroxiredoxin Prdx1 in erythrocyte antioxidant defence and tumour suppression. *Nature* **424**, 561–565 (2003).
- Snijders, A. M. et al. Shaping of tumour and drug-resistant genomes by instability and selection. *Oncogene* **22**, 4370–4379 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank G. Stark for advice with the PALA assay; M. LaBarge, M. Adriance and other members of the Bissell laboratory for discussions; J. Campisi for critical reading of the manuscript; and C. Chen for technical assistance. This work was supported by grants from the OBER office of the Department of Energy and an Innovator award from the Department of Defense (to M.J.B.) and from the National Institutes of Health (to M.J.B. and Z.W.), and by fellowships from the American Cancer Society (D.C.R.), the National Cancer Institute (L.E.L.), the Department of Defense (H.L. and C.M.N.) and the California Breast Cancer Research Program (J.E.F.).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.J.B. (mjbissell@lbl.gov) or D.C.R. (dcradisky@lbl.gov).

LETTERS

Kinase-regulated quantal assemblies and kiss-and-run recycling of caveolae

Lucas Pelkmans¹ & Marino Zerial¹

A functional genomics approach has revealed that caveolae/raft-mediated endocytosis is subject to regulation by a large number of kinases¹. Here we explore the role of some of these kinases in caveolae dynamics. We discover that caveolae operate using principles different from classical membrane trafficking. First, each caveolar coat contains a set number (one 'quantum') of caveolin-1 molecules. Second, caveolae are either stored as in stationary multi-caveolar structures at the plasma membrane, or undergo continuous cycles of fission and fusion with the plasma membrane in a small volume beneath the surface, without disassembling the caveolar coat. Third, a switch mechanism shifts caveolae from this localized cycle to long-range cytoplasmic transport. We have identified six kinases that regulate different steps of the caveolar cycle. Our observations reveal new principles in caveolae trafficking and suggest that the dynamic properties of caveolae and their transport competence are regulated by different kinases operating at several levels.

Caveolae are involved in dynamic cellular processes such as signal transduction² and endocytosis³, but appear to be constitutively immobile^{4,5}. Inside the cell, caveolar vesicles dock on, fuse with and are released from endocytic organelles without disassembling the caveolin-1 (CAV1) coat⁶. It is not known whether caveolae are assembled *de novo* at the cell surface or originate from intracellular pools, whether they interact dynamically with the plasma membrane, how they participate in membrane trafficking and how this is regulated by kinases¹. We address these questions using a combination of quantitative total internal reflection fluorescence microscopy (TIR-FM)⁷, computational analysis and functional genomics using RNA interference (RNAi)¹.

Caveolae were visualized in HeLa cells stably expressing CAV1 labelled with green fluorescent protein (GFP)⁶ (see Supplementary Information for detailed methods). Fitting the intensity distributions of the visualized spots to multiple gaussian curves revealed that caveolar structures were assembled from unit-sized building blocks, with a constant fluorescence intensity quantum (q). (Fig. 1a, b). Quantal incorporation into synaptic vesicles has been shown for neurotransmitters and lipid dyes^{8,9}. To assign each gaussian peak to a morphologically known structure, we estimated their size. CAV1-GFP spot sizes of $q_{n<4}$ were as large as the diffraction limit of our imaging system (~ 200 nm) (Supplementary Fig. S1a–c), but size differences became apparent for higher q values ($q \geq 4$). Together with previous ultrastructural analysis of CAV1-GFP on the cell surface^{5,10}, we conclude that the q_1 structures (~ 50 nm in size) are individual caveolae, and that $q_{>1}$ structures are grape-like clusters of multiple caveolae.

Using mouse *Cav1*^{-/-} fibroblasts¹¹ expressing CAV1-GFP, the number of CAV1 molecules in q_1 structures was estimated to be 144 ± 39 (Supplementary Fig. S1). The intensity of CAV1-GFP q_1 quanta in HeLa cells was ~ 0.56 of that in *Cav1*^{-/-} fibroblasts (Fig. 1b and Supplementary Fig. S1f). As CAV1-GFP is expressed at similar

levels as endogenous CAV1 in HeLa cells⁶, this indicates that quanta of both cell lines contain a similar number of CAV1 molecules, and that the GFP moiety does not disturb incorporation or quantal assembly. The number of CAV1 molecules in quanta was independent of expression level (Supplementary Fig. S1d, f). However, upon 5–10-fold overexpression of CAV1-GFP, CAV1 also appeared in surface structures at sub-quantal intensity (0.1 – $0.3q$; Supplementary Fig. S1e, f) that, in contrast to individual caveolae, showed lateral mobility within the plane of the membrane (Supplementary Video 1). These might represent CAV1 oligomers that failed to be incorporated into caveolae owing to saturation of the assembly mechanism.

We observed that many caveolae are not fixed at the plasma membrane, but instead appear and disappear (Supplementary Videos 2, 2a), as indicated by rapid quantal increases and complete losses of fluorescence intensity (Fig. 1c, d and Supplementary Videos 3, 3a, 3b). During docking, spot intensity remained constant and CAV1-GFP did not diffuse into the surrounding membrane (Supplementary Fig. S1i), suggesting that pre-assembled caveolar vesicles keep their structural integrity while cycling between the cytoplasm and the cell surface⁶.

Most structures showing kiss-and-run interactions (90/398) were individual caveolae (Fig. 1d). In fact, over a period of 300 s, 45% (71/158) of all individual caveolae were dynamic. Occasionally, we observed the docking of a q_1 vesicle and the subsequent docking of a second q_1 vesicle, resulting in the formation of a q_2 structure (Supplementary Fig. S1j). We thus conclude that multi-caveolar assemblies are formed by multiple individual caveolar vesicles docking on top of one another. These are stably attached to the cell surface, whereas half of the individual caveolae show continuous kiss-and-run interactions with the plasma membrane.

To demonstrate that caveolae can undergo a cycle of docking, fusion and fission, we used pH quenching¹² with dual colour TIR-FM. The outside of caveolar vesicles was visualized with CAV1 tagged with monomeric red fluorescent protein (CAV1-mRFP) and the vesicle lumen was visualized with fluorescein-isothiocyanate-labelled Cholera toxin B (FITC-ChTxB)¹³. Because the interior of caveolar vesicles and caveosomes is pH neutral¹⁴, we used rapid loss of FITC signal upon acidification as the readout for fusion.

Using TIR-FM, we found that most caveolar structures incorporated the FITC-ChTxB toxin, including those showing kiss-and-run interactions with the plasma membrane (Supplementary Fig. S2a and Supplementary Video 4). Upon acidification of the extracellular medium to pH 5.5, the FITC signal was quenched, predominantly in multi-caveolar assemblies; it was not quenched in several individual caveolae, suggesting that these caveolae were docked on, but not continuous with the plasma membrane (Fig. 1e). However, the FITC signal was rapidly quenched to background levels in these structures about 1 s after docking (Fig. 1f, g, Supplementary Fig. S2b and Supplementary Videos 4a–c). Similar quenching kinetics were observed at an extracellular pH of 6.2, a pH at which some signal

¹Max Planck Institute for Molecular Cell Biology and Genetics, Pfötenhauerstrasse 108, 01307 Dresden, Germany.

remained (in agreement with the pK_a of FITC) (Fig. 1g). Out of 32 docking vesicles, 27 became rapidly quenched in a *N*-ethylmaleimide (NEM)-sensitive manner (Supplementary Fig. S2b, d). Addition of ionophores, which equilibrate protons across membranes, resulted in quenched FITC signal in docking caveolar vesicles (Supplementary Fig. S2b). Together, these observations indicate that intracellular caveolar vesicles that have incorporated ChTxB during a previous visit to the cell surface or to intracellular, pH-neutral organelles like caveosomes^{6,14}, can dock at and fuse with the plasma membrane. Alternatively, model membrane experiments suggest that exposure of

the vesicle lumen to the extracellular medium during kiss-and-run interactions might be determined by length-controlled accessibility of a tubular membranous neck continuously connected to the surface¹⁵.

Under conditions of neutral extracellular pH, FITC-ChTxB was not released into the surrounding membrane after fusion of a caveolar vesicle. However, release did occur when the glycosphingolipid-bound toxin was exocytosed by other, CAV1-mRFP-negative vesicles (Fig. 1f and Supplementary Fig. S2c), in a manner similar to the release of lipid dyes or transmembrane proteins from secretory

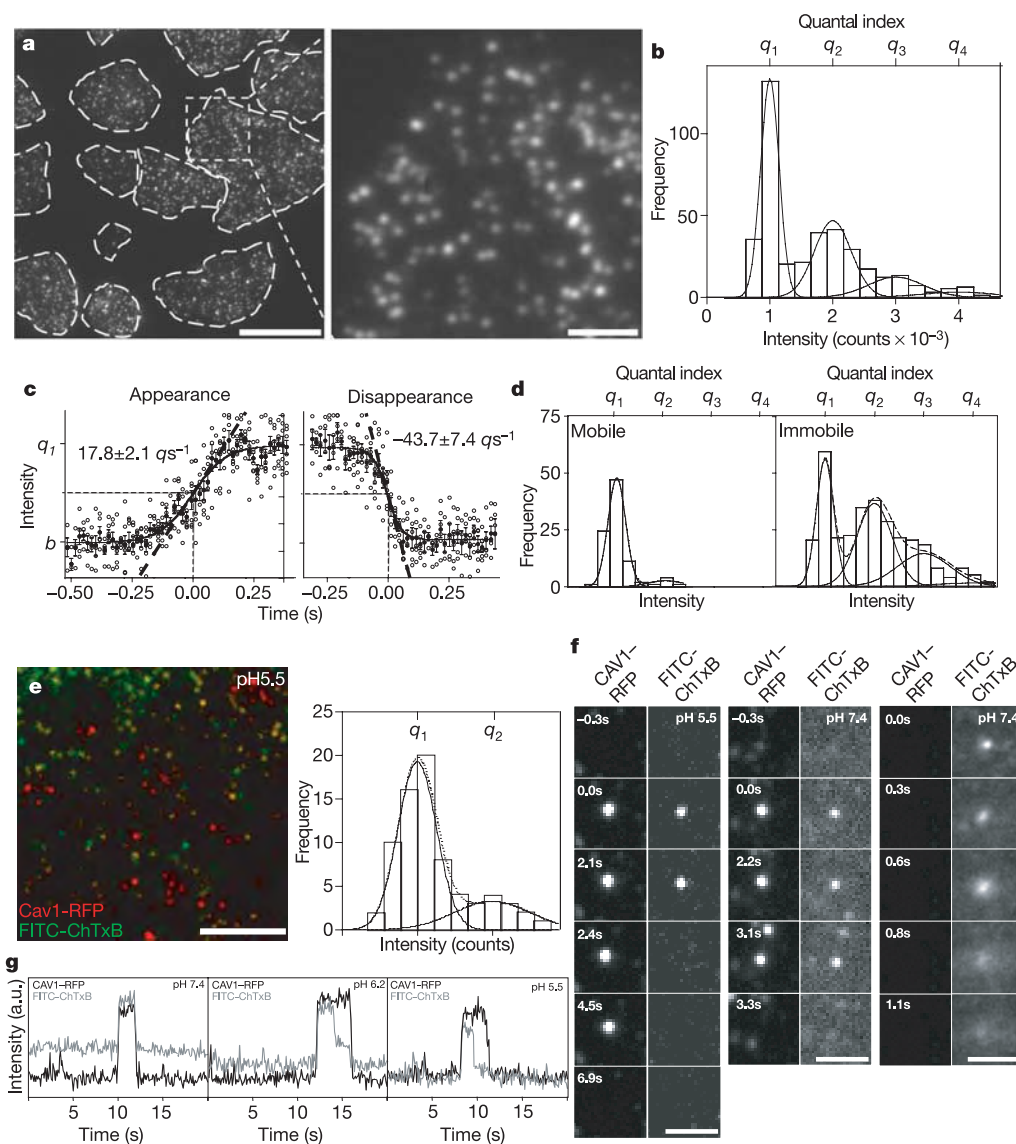


Figure 1 | Quantal assembly and kiss-and-run of caveolar structures.

a, TIR-FM of HeLa cells stably expressing CAV1-GFP. Scale bars, 10 μ m (left) and 2 μ m (right). **b**, Intensity histogram of 396 CAV1-GFP-positive structures in HeLa cells, fitted to 4 gaussian curves (degrees of freedom (d.f.) = 13; $R^2 = 0.99$; absolute sum of squares (S.S.) = 154.8; residuals ($S_{y,x}$) = 3.45). The mean of the first peak is defined as q_1 . **c**, Plots of intensity over time (17-ms interval) of 12 appearing and 10 disappearing structures in TIR-FM, and sigmoidal curve fits (solid lines) of average ($\pm$ s.e.m.) values (filled circles) (appearing: d.f. = 344, $R^2 = 0.82$, S.S. = 1.7, $S_{y,x} = 0.07$; disappearing: d.f. = 233, $R^2 = 0.81$, S.S. = 1.3, $S_{y,x} = 0.07$). The rates of appearance ($17.8 \pm 2.1 q s^{-1}$) and disappearance ($-43.7 \pm 7.4 q s^{-1}$) and the level of background signal (b) are shown on the graph. **d**, Intensity histograms and gaussian fits of 90 docking/mobile (d.f. = 6, $R^2 = 1.00$, S.S. = 4.8, $S_{y,x} = 0.89$) and 308 fixed/immobile structures (d.f. = 14,

$R^2 = 0.97$, S.S. = 146.3, $S_{y,x} = 3.23$). Of the 90 mobile spots, 71 are q_1 and 7 are q_2 . Of the 308 fixed spots, 87 are q_1 and 221 are $q_{n>1}$. **e**, Dual-colour TIR-FM, intensity histogram and gaussian fit of 73 caveolar structures with non-quenchable FITC-ChTxB (d.f. = 10, $R^2 = 0.97$, S.S. = 12.9, $S_{y,x} = 1.14$). Scale bar, 5 μ m. **f**, Selected frames from 100-ms interval recordings. Left, FITC-ChTxB in a caveolar vesicle, quenched by low extracellular pH during docking. Middle, FITC-ChTxB in a caveolar vesicle remains sequestered at neutral pH during docking. Right, FITC-ChTxB in a non-caveolar vesicle is released into the surrounding membrane at neutral pH during docking. Scale bar, 1 μ m. **g**, CAV1-RFP and FITC-ChTxB fluorescence intensity traces (100-ms intervals; no background correction) during kiss-and-run of an individual caveola at extracellular pH 7.4, 6.2 or 5.5. Fusion is detected as rapid quenching of the FITC signal before the RFP signal disappears.

vesicles¹⁶. Thus, during the fusion of individual caveolae with the plasma membrane, the caveolar coat does not fully dissociate, and it retains the ability to sequester a ganglioside GM1-associated multivalent ligand, similar to intracellular caveolar vesicles⁶.

To follow the origin and fate of the cycling population of caveolar vesicles, we combined TIR-FM with epifluorescence microscopy (Epi-FM)¹⁶. A spot appearing in TIR-FM was invariably observed first in Epi-FM (Fig. 2a, c and Supplementary Videos 5, 5a), confirming that caveolae are not assembled *de novo* at the surface, but instead are pre-assembled in intracellular caveolar vesicle stores and then transported to the cell surface. Indeed, we could directly monitor a FITC-ChTx-B-containing, pre-existing vesicle appearing from the cytoplasm, docking on the surface, exposing its lumen to the extracellular space, and internalizing again (Supplementary Fig. S3). Between docking events, we occasionally observed rapid, directional movement. The majority of mobile caveolae, however, repeatedly docked in a small area of the plasma membrane, and combined this with short-range motility below the surface (Fig. 2b and Supplementary Video 5b). We performed statistical analysis of the spatial distribution of all docking events over a 300-s recording

period, and confirmed that most were confined to an area of 3–8 μm^2 (Fig. 2d). Apparently individual caveolae undergo (1) continuous short-range cycles of fusion and internalization, during which they do not travel long distances but are restricted within a small cytoplasmic volume beneath the cell surface, and (2) occasional trafficking between the plasma membrane and intracellular pools. This is consistent with a slow exchange between these two pools and a seemingly fixed position of caveolar structures on the cell surface^{4,5}.

We next performed an in-depth analysis of six kinases identified as potential candidate regulators of caveolae/raft-mediated endocytosis that are also involved in infectious entry of simian virus 40 (SV40)¹. A series of RNAi phenotypes reflected functions at distinct steps of the caveolar cycle described above. First, the silencing of *ARAF1* (also known as *ARAF*), a serine/threonine kinase involved in mitogenic signalling, resulted in diffuse CAV1-GFP staining that was laterally mobile, in addition to the characteristic spot-like pattern (Fig. 3a, Supplementary Fig. S2c and Supplementary Video 6). We suggest that in the absence of *ARAF1*, the caveolar coat is less stable or inefficiently assembled.

Second, ablation of *MGC26597* (a hypothetical kinase with a

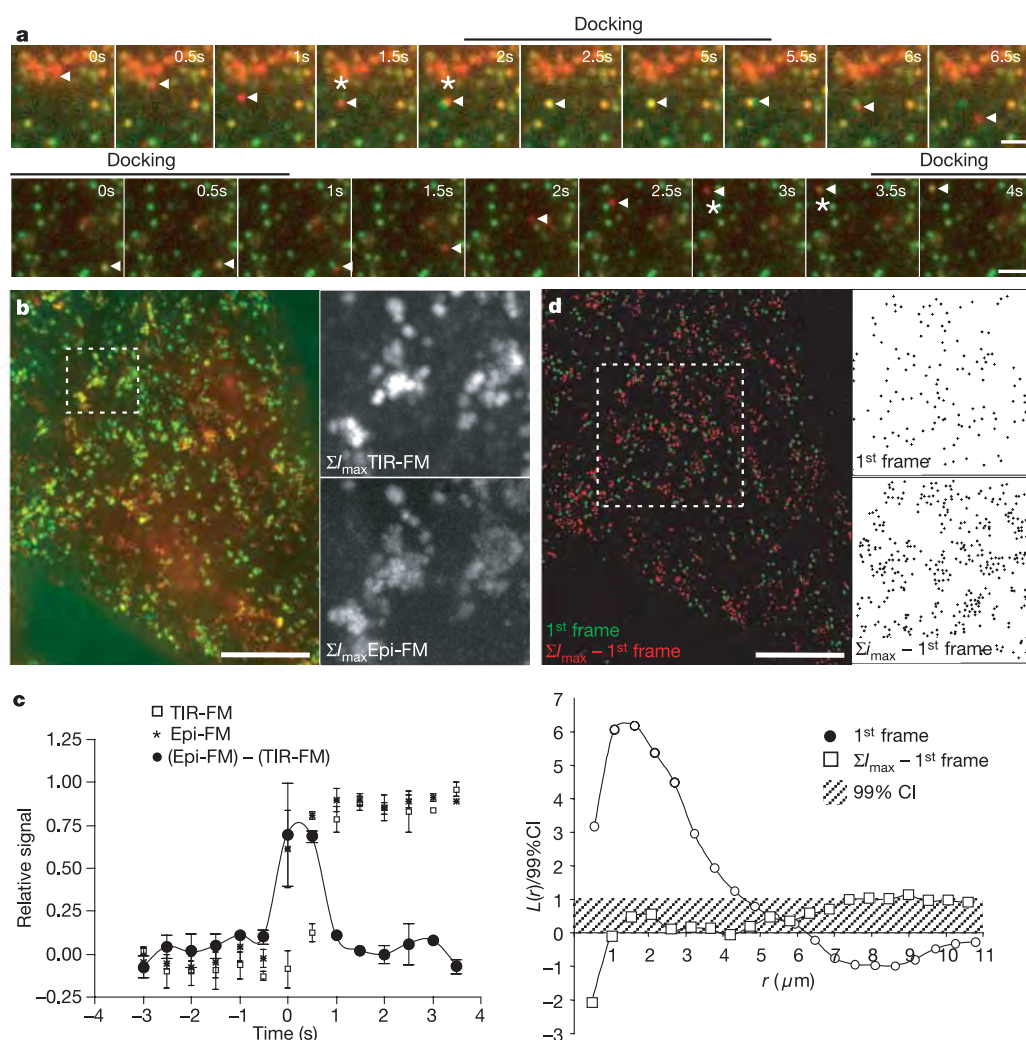


Figure 2 | Cycling is restricted to local areas beneath the plasma membrane. **a**, Combined TIR-FM (green) and Epi-FM (red). Plasma membrane kiss-and-run by a caveolar vesicle departing from an intracellular organelle (top) and rapid, directional transport of a caveolar vesicle between two docking events (bottom). Time is relative to the first frame. Scale bars, 1 μm (top), 2 μm (bottom). **b**, Maximum intensity projections (ΣI_{max}) of TIR-FM (green) and Epi-FM (red) time-lapse sequence reveal areas of clustered spots that co-localize with each other. Scale bar, 10 μm . **c**, Average

($\pm$ s.e.m.) relative intensities of docking sites measured by TIR-FM (open squares) and Epi-FM (stars). The Epi-FM signal appears 1–2 frames (500–1,000 ms) before the TIR-FM signal (see asterisks in **a**). **d**, Quantification of clustering events by coordinate determination and Ripley's *K* function analysis of caveolar structures on the surface (at the first frame, green) and of docking events during time-lapse (ΣI_{max} minus the first frame, red). Scale bar, 10 μm .

phosphoinositide 4-phosphate 5-kinase (PI(4)P5K) homology domain) or *SRC* resulted in clustering of caveolae (Fig. 3a), leading to an increase in multi-caveolar assemblies ($q_{n>3}$) at the expense of individual caveolae. As a result, caveolar structures accumulated at the cell surface, and both the dynamics of caveolae (Fig. 3b, c) and the internalization of SV40 particles were strongly reduced (Supplementary Fig. S2d and Supplementary Video 7).

Third, silencing of two serine/threonine kinases (*KIAA0999* and *MAP3K2*), strongly reduced kiss-and-run dynamics, leading to an accumulation of caveolar structures at the cell surface (Fig. 3b, c and Supplementary Video 8) without affecting caveolae coat assembly or clustering (not shown). Conversely, knockdown of the Ser/Thr MAPK-related kinase *DYRK3* strongly increased the dynamics of

caveolar vesicles (Fig. 3b, c) and had a minor destabilizing effect on the caveolar coat (Supplementary Video 9).

Increased dynamics were also observed upon exposure to okadaic acid or SV40, both known to stimulate caveolae-mediated endocytosis^{4,17} (Fig. 3b, c and Supplementary Video 10). There was no net change in the number of caveolae at the cell surface, but we did detect a three- to fivefold increase in the number of docking events, which occurred at random positions on the plasma membrane (Fig. 3c, d). Random docking is also seen after actin depolymerization (data not shown), and this is initiated by SV40 (ref. 4). Because cell-surface and intracellular pools of CAV1 exchange upon addition of okadaic acid⁵ or SV40 (data not shown), we propose that stimulation of caveolae-mediated endocytosis requires elimination

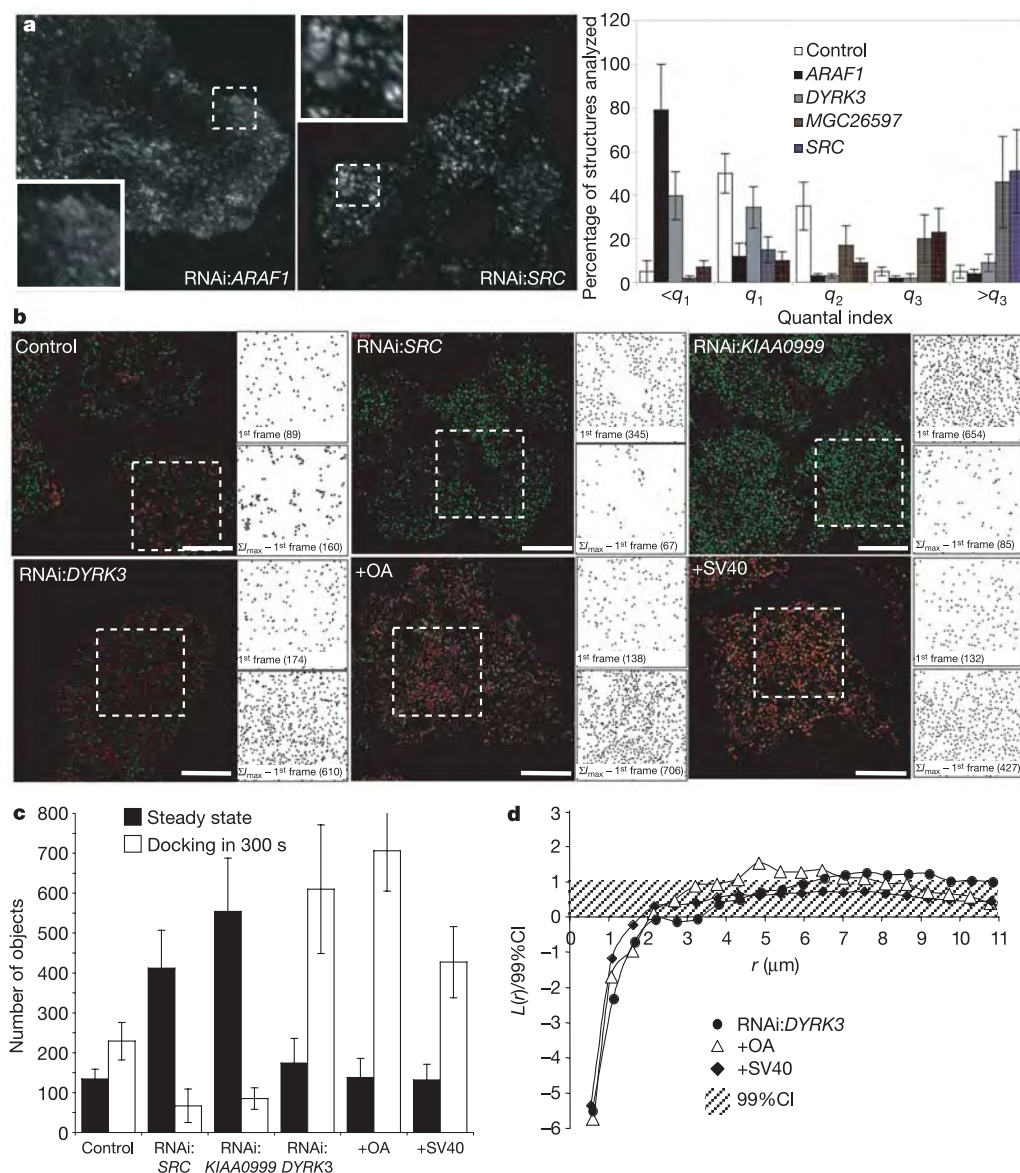


Figure 3 | Caveolae coat stability, clustering and cycling mode are controlled by kinases. **a**, TIR-FM and quantal index analysis (average values $\pm$ s.e.m.; $n = 212$ –479; 3 cells) of CAV1-GFP shows that RNAi-mediated knockdown of *ARAF1* (RNAi:ARAF1) results in diffuse staining ($<q_1$) and RNAi:SRC results in the formation of large, multi-caveolar clusters. **b**, Coordinate determination of caveolar docking sites during RNAi-mediated knockdown of *SRC*, *KIAA0999* or *DYRK3*, and after treatment with 1 μM okadaic acid (OA) or exposure to SV40. Scale bars, 10 μm . **c**, Average CAV1-GFP spot density ($\pm$ s.e.m.) determined for the first

frame and ΣI_{max} minus the first frame, from 8 areas of 21.6 μm^2 each. RNAi:SRC and RNAi:KIAA0999 reduce the number of docking events and increase the number of caveolar structures on the cell surface. RNAi:DYRK3, OA-treatment and SV40 incubation increase the number of docking events. **d**, Quantification of clustering of docking events after RNAi:DYRK3, OA-treatment and SV40 exposure shows loss of clustering. Regions of the curve above the 99% CI indicate clusters of caveolae within an area size of radius r (see Methods).

of the spatial restrictions on caveolae dynamics imposed by the actin cytoskeleton⁴, resulting in a switch from futile kiss-and-run cycles to long-range, microtubule-dependent transport cycles. This results in 'committed' internalization of caveolae to intracellular organelles and their replenishment with caveolar vesicles from intracellular locations.

In this study, we have identified new principles of caveolae membrane trafficking. First, a quantal number of 144 ± 39 molecules of CAV1 are incorporated in one caveolar coat. Second, whereas caveolae not engaged in trafficking are stored as clusters in multi-caveolar assemblies that communicate with the extracellular space, transport-competent caveolae undergo constitutive kiss-and-run cycles in a small volume underneath the plasma membrane, during which the caveolar coat stays intact and sequesters multivalent, sphingolipid-bound cargo. Third, a molecular switch changes the mode of cycling from short- to long-range and activates caveolae-mediated transport to intracellular locations (such as caveosomes¹⁴ and, to a lesser extent, early endosomes⁶). Intracellular caveolar vesicles are concomitantly recycled to the plasma membrane.

We have identified six kinases that regulate coat stability (*ARAF1*), caveolae clustering (*SRC*, *MGC26597*), kiss-and-run dynamics (*KIAA0999*, *MAP3K2*) and long-range cycling (*DYRK3*). Mitogenic signalling downstream of *ARAF1* might act through (de)stabilizing the caveolar signalling scaffold² and differentially shaping CAV1-GFP-positive membranes¹.

Ablation of *SRC* or *MGC26597* causes the same caveolae-clustering phenotype as overexpression of the GTP hydrolysis-deficient mutant of dynamin2 (ref. 18; Supplementary Fig. S2e), and blocks infectious entry of SV40. This is consistent with suggested roles for *SRC* in caveolae function and/or dynamics^{19,20}. The specific recruitment and activation of dynamin2 on caveolar structures might be regulated by *MGC26597* (through phosphatidylinositol-4,5-bisphosphate (PtdIns(4,5)P₂) synthesis; ref. 21) and *SRC*²². *MGC26597* could be a caveolae/lipid raft-mediated endocytosis-specific PI(4)P5K, as it is not required for clathrin-mediated endocytosis¹.

Differential regulation of short- and long-distance cycling between the cell surface and intracellular organelles by *KIAA0999*, *MAP3K2* and *DYRK3* points to the existence of kinase-regulated molecular machinery that can activate these two modes of trafficking and switch between them. This probably involves changes in the cortical actin cytoskeleton and activation of microtubule-dependent motility^{1,10,23}.

Finally, we can make suggestions regarding different modes of caveolae- and raft-dependent endocytosis. First, a ligand-activated switch from short- to long-range cycles induces committed internalization of caveolae. Second, directional transport independent of this switch might be established when locally cycling caveolae transiently fuse with nearby organelles when a cue induces ligand release⁶. Third, a mechanism might be in place to destabilize the caveolar coat at the plasma membrane, allowing release of lipid rafts and their internalization by means of non-caveolae-derived vesicles, which could be under the control of mitogenic kinases¹. Continuing with the combined approaches of functional genomics and TIR-FM will allow systems analysis and a more detailed understanding of these related processes.

METHODS

For full details of the methods used, please refer to the Supplementary Information.

Cell culture and transfection. All cells were cultured in medium containing 10% fetal calf serum. The medium for cells expressing CAV1-GFP or CAV1-mRFP contained $500 \mu\text{g ml}^{-1}$ G418. Transient transfections were performed by lipofection using FuGene6 (Roche) and short interfering RNA was transfected using Oligofectamine (Invitrogen). For imaging purposes, cells were plated in glass-bottomed micro-well dishes (MatTek) and maintained in CO₂-independent medium (GibcoBRL). Time-lapse experiments were performed at 25 °C and 37 °C.

Imaging. Imaging was performed using an Olympus IX70 inverted microscope equipped with a dual-port condenser (TILL Photonics), an argon-krypton laser

(Innova 70C Spectrum, Coherent) and a 100 W mercury arc lamp light source, to allow both Epi-FM and TIR-FM. The laser beam was focused at an off-axis position in the back focal plane of high numerical aperture $\times 63/\text{NA } 1.45$ (Olympus) or $\times 100/\text{NA } 1.45$ (Zeiss) oil immersion objectives, such that the laser beam struck the interface between the glass and cell at an angle less than 55°. As a result, the laser light underwent total internal reflection, leading to the excitation of molecules within 100 nm above the interface only. For dual-colour experiments, a beam splitter (Dual-View Micro-Imager, Optical Insights) was used to project green and red components side by side onto a back-illuminated CCD camera (Micromax 512BFT, Roper Scientific). Time-lapse sequences were acquired at 100–500-ms intervals. For high-speed imaging (17-ms intervals), a 50×50 pixel region of interest was selected.

Identification of quantal assembly. Background-corrected intensity histograms were fitted to functions consisting of the sum of 2 or 4 gaussian curves:

$$F(I) = \sum_{k=1}^{k=n} v_k e^{-\frac{(I-q_k)^2}{2\sigma_k^2}} \quad (1)$$

where F is the number of caveolar structures at a given intensity I ; v_k is the amplitude of the k th peak, σ_k is the variance of the intensity of an individual quantum and q is the intensity of an individual quantum. Structures with I values with >95% probability of belonging to the $k = 1$ gaussian curve (q_1 spots) were used to determine the number of CAV1-GFP molecules.

Determination of the number of incorporated caveolin-1 molecules. To calculate the number of CAV1-GFP molecules in q_1 spots, we compared the distribution of integrated gaussian curves fitted to radial sweeps of 375 rotavirus-like particles (VLPs) containing 120 molecules of GFP with the distribution of integrated gaussian curves fitted to radial sweeps of 405 q_1 spots (Supplementary Fig. S1).

Analysis of fusion events. HeLa cells expressing CAV1-mRFP were allowed to bind $\sim 10,000$ molecules of FITC-ChTxB for 5 min at 37 °C and were then extensively washed with warm CO₂-independent medium. Cells were transferred to the microscope and the imaging system was prepared to record kiss-and-run dynamics. An equal amount of citric acid-buffered, CO₂-independent medium was subsequently added to obtain a final pH of 5.5 (or 6.2) and dynamics of CAV1-mRFP and FITC-ChTxB were immediately recorded using dual-colour TIR-FM. Each video was finished within 1 min of acidification of the medium.

Statistical analysis using Ripley's K function. Coordinates of docking events on the plasma membrane were used to determine the distance between each docking event and to analyse the second-order property of the point pattern, defined by Ripley's K function:

$$K(r) = \frac{N(r)}{\lambda} \quad (2)$$

where $N(r)$ is the expected number of neighbours within a distance r , and $K(r)$ is $N(r)$ normalized to λ , the density of the pattern. The function used in our analysis is derived from equation (2) and describes $L(r)$, which is $K(r)$ corrected for values at complete spatial randomness (CSR) at which $N(r)$ is $\lambda\pi r^2$:

$$L(r) = \sqrt{\frac{K(r)}{\pi}} - r \quad (3)$$

At CSR, which is used as the benchmark in our analysis, $L(r)$ is 0. Monte Carlo simulations were performed to calculate the 99% confidence interval (CI) for CSR and $L(r)$ values are presented relative to the 99%CI. $L(r)$ functions were plotted and regions of the curve that emerge above the 99%CI indicate significant clustering within an area size of radius r .

Received 1 March; accepted 25 May 2005.

1. Pelkmans, L. et al. Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis. *Nature* doi:10.1038/nature03571 (this issue).
2. Razani, B., Woodman, S. E. & Lisanti, M. P. Caveolae: from cell biology to animal physiology. *Pharmacol. Rev.* **54**, 431–467 (2002).
3. Pelkmans, L. & Helenius, A. Endocytosis via caveolae. *Traffic* **3**, 311–320 (2002).
4. Pelkmans, L., Puntener, D. & Helenius, A. Local actin polymerization and dynamin recruitment in SV40-induced internalization of caveolae. *Science* **296**, 535–539 (2002).
5. Thomsen, P., Roepstorff, K., Stahlhut, M. & van Deurs, B. Caveolae are highly immobile plasma membrane microdomains, which are not involved in constitutive endocytic trafficking. *Mol. Biol. Cell* **13**, 238–250 (2002).
6. Pelkmans, L., Burli, T., Zerial, M. & Helenius, A. Caveolin-stabilized membrane domains as multifunctional transport and sorting devices in endocytic membrane traffic. *Cell* **118**, 767–780 (2004).

7. Axelrod, D., Burghardt, T. P. & Thompson, N. L. Total internal reflection fluorescence. *Annu. Rev. Biophys. Bioeng.* **13**, 247–268 (1984).
 8. Katz, B. Quantal mechanism of neural transmitter release. *Science* **173**, 123–126 (1971).
 9. Ryan, T. A., Reuter, H. & Smith, S. J. Optical detection of a quantal presynaptic membrane turnover. *Nature* **388**, 478–482 (1997).
 10. Mundy, D. I., Machleidt, T., Ying, Y. S., Anderson, R. G. & Bloom, G. S. Dual control of caveolar membrane traffic by microtubules and the actin cytoskeleton. *J. Cell Sci.* **115**, 4327–4339 (2002).
 11. Drab, M. *et al.* Loss of caveolae, vascular dysfunction, and pulmonary defects in caveolin-1 gene-disrupted mice. *Science* **293**, 2449–2452 (2001).
 12. Miesenböck, G., De Angelis, D. A. & Rothman, J. E. Visualizing secretion and synaptic transmission with pH-sensitive green fluorescent proteins. *Nature* **394**, 192–195 (1998).
 13. Parton, R. G. Ultrastructural localization of gangliosides; GM1 is concentrated in caveolae. *J. Histochem. Cytochem.* **42**, 155–166 (1994).
 14. Pelkmans, L., Kartenbeck, J. & Helenius, A. Caveolar endocytosis of simian virus 40 reveals a new two-step vesicular-transport pathway to the ER. *Nature Cell Biol.* **3**, 473–483 (2001).
 15. Frolov, V. A., Lizunov, V. A., Dunina-Barkovskaya, A. Y., Samsonov, A. V. & Zimmerberg, J. Shape bistability of a membrane neck: a toggle switch to control vesicle content release. *Proc. Natl Acad. Sci. USA* **100**, 8698–8703 (2003).
 16. Toomre, D., Steyer, J. A., Keller, P., Almers, W. & Simons, K. Fusion of constitutive membrane traffic with the cell surface observed by evanescent wave microscopy. *J. Cell Biol.* **149**, 33–40 (2000).
 17. Parton, R. G., Joggerst, B. & Simons, K. Regulated internalization of caveolae. *J. Cell Biol.* **127**, 1199–1215 (1994).
 18. Henley, J. R., Krueger, E. W., Oswald, B. J. & McNiven, M. A. Dynamin-mediated internalization of caveolae. *J. Cell Biol.* **141**, 85–99 (1998).
 19. Minshall, R. D. *et al.* Endothelial cell-surface gp60 activates vesicle formation and trafficking via G_i-coupled Src kinase signalling pathway. *J. Cell Biol.* **150**, 1057–1070 (2000).
 20. Sharma, D. K. *et al.* Selective stimulation of caveolar endocytosis by glycosphingolipids and cholesterol. *Mol. Biol. Cell* **15**, 3114–3122 (2004).
 21. Salim, K. *et al.* Distinct specificity in the recognition of phosphoinositides by the pleckstrin homology domains of dynamin and Bruton's tyrosine kinase. *EMBO J.* **15**, 6241–6250 (1996).
 22. Shajahan, A. N. *et al.* Role of Src-induced dynamin-2 phosphorylation in caveolae-mediated endocytosis in endothelial cells. *J. Biol. Chem.* **279**, 20392–20400 (2004).
 23. Pelkmans, L. & Helenius, A. Insider information: What viruses tell us about endocytosis. *Curr. Opin. Cell Biol.* **15**, 414–422 (2003).
- Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.
- Acknowledgements** We thank K. Anderson and S. Diez for help with TIR-FM, and M. Drab for *Cav1*^{−/−} fibroblasts. D. Dorris, R. Günther, I. Baines and the Max Planck Institute for Molecular Cell Biology and Genetics are acknowledged for having made possible the kinome screen of endocytosis. We thank D. Meder, K. Simons, A. Helenius and Y. Kalaidzidis for discussions and critical reading of the manuscript. L.P. would like to thank A. Helenius for support. This work was supported by grants from The Max Planck Society 'RNAi interference' initiative and the Bunderministerium für Bildung und Forschung. L.P. is a Marie Curie Fellow.
- Author Information** Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.P. (pelkmans@mpi-cbg.de).

LETTERS

Molecular basis of photoprotection and control of photosynthetic light-harvesting

Andrew A. Pascal¹, Zhenfeng Liu², Koen Broess³, Bart van Oort³, Herbert van Amerongen³, Chao Wang², Peter Horton⁴, Bruno Robert¹, Wenrui Chang² & Alexander Ruban⁴

In order to maximize their use of light energy in photosynthesis, plants have molecules that act as light-harvesting antennae, which collect light quanta and deliver them to the reaction centres, where energy conversion into a chemical form takes place. The functioning of the antenna responds to the extreme changes in the intensity of sunlight encountered in nature^{1–3}. In shade, light is efficiently harvested in photosynthesis. However, in full sunlight, much of the energy absorbed is not needed and there are vitally important switches to specific antenna states, which safely dissipate the excess energy as heat^{2,3}. This is essential for plant survival⁴, because it provides protection against the potential photo-damage of the photosynthetic membrane⁵. But whereas the features that establish high photosynthetic efficiency have been highlighted⁶, almost nothing is known about the molecular nature of the dissipative states. Recently, the atomic structure of the major plant light-harvesting antenna protein, LHCII, has been determined by X-ray crystallography⁷. Here we demonstrate that this is the structure of a dissipative state of LHCII. We present a spectroscopic analysis of this crystal form, and identify the specific changes in configuration of its pigment population that give LHCII the intrinsic capability to regulate energy flow. This provides a molecular basis for understanding the control of photosynthetic light-harvesting.

The reversible switch between these two antenna states of energy harvesting and energy dissipation has been well-characterized physiologically⁸. It is known as non-photochemical quenching (NPQ), because, by decreasing the excitation level, it reduces the yield of (or quenches) chlorophyll fluorescence. NPQ covers a range of responses operating on different timescales and with different strengths of quenching, which tune the antenna to the prevailing light conditions⁹. The antenna is composed of specialized membrane-bound light-harvesting pigment–protein complexes, in which chlorophylls and carotenoids are organized in a very ordered manner at significant density. The precise molecular mechanisms through which the antenna could reversibly switch between fundamentally different states remain controversial^{10–12}. In particular, there is little understanding of how pigment function could be altered within these complexes so as to form efficient energy quenchers.

When removed from the photosynthetic membrane, the main trimeric light-harvesting antenna complex, LHCII, is highly fluorescent, indicating a low rate of energy dissipation. The lifetime of the excited state is typically around 4 ns (ref. 13; see also below). However, when the complexes self-associate into oligomers or aggregates, the fluorescence is highly quenched¹⁴, with a range of lifetimes between 0.2 and 1.5 ns (refs 13, 15). The increase in energy dissipation within LHCII oligomers compared to trimers provides a

model for understanding *in vivo* NPQ¹⁶, because of the strong similarities between the two quenching processes². In the atomic structure of LHCII, derived from X-ray crystallography⁷, the contacts between adjacent trimers within the proteoliposome vesicle are minimal, with only two pairs of chlorophylls in weak van der Waals interaction. No transmembrane domains are involved in the contact (Fig. 1). Therefore, each trimer in the crystal is almost functionally separate. We were therefore surprised to find that the fluorescence of LHCII in these crystals is quenched (Fig. 2). To measure quenching, we have used FLIM (fluorescence lifetime imaging microscopy), a form of laser-scanning microscopy that excites chromophores only within single pixels. For each pixel, a

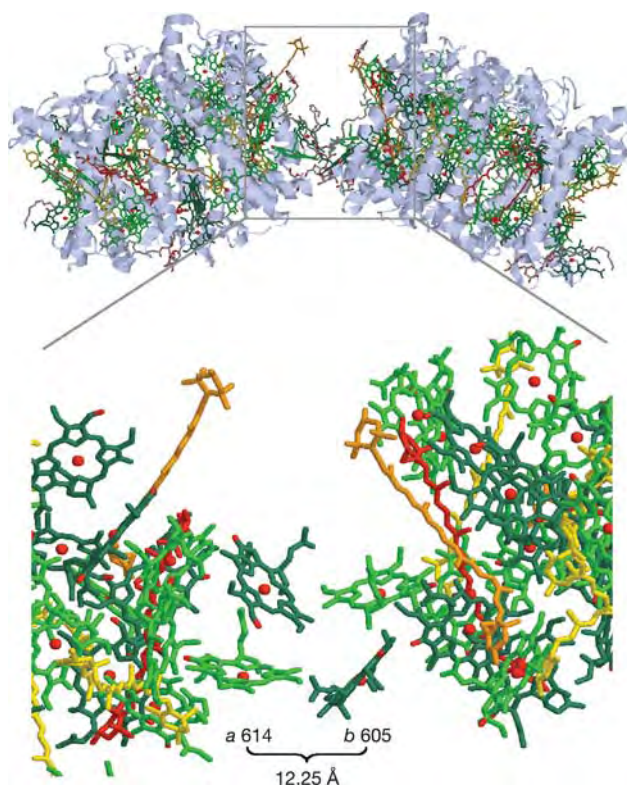


Figure 1 | Crystal structure of LHCII. Two interacting trimers are shown from the icosahedral vesicle in the crystal⁷. The enlargement shows the two pairs of their closest peripheral pigments, Chl a 614 and Chl b 605.

¹Service de Biophysique des Fonctions Membranaires, URA2096/CNRS and DBJC/CEA, CEA-Saclay, 91191 Gif-sur-Yvette Cedex, France. ²National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, 15 Datun Road, Chaoyang District, Beijing 100101, China. ³Laboratory of Biophysics, Wageningen University, PO Box 8128, 6700 ET, Wageningen, The Netherlands. ⁴Department of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield S10 2TN, UK.

fluorescence decay curve is measured by time-correlated single photon counting. Fluorescence lifetimes are independent of chromophore concentration and scatter, thus readily allowing comparison of the fluorescence lifetimes and monitoring of the sample homogeneity. The fluorescence lifetimes of each pixel are displayed to give the false colour image in Fig. 2b. The homogeneity of the image indicates that the lifetime was remarkably uniform and that there is only very minor variation within the crystal (Fig. 2b). The average fluorescence decays from the crystal fit a single exponential and indicate a lifetime of 0.89 ns (Fig. 2c), resembling the oligomeric rather than the trimeric state, indicating the presence of energy traps not present in the trimer. (Using the same apparatus, the lifetime of LHCII oligomers was found to be 0.65 ns.)

In order to identify the molecular basis of the change in excitation lifetime, a spectroscopic analysis of the LHCII crystal was carried out. The fluorescence emission spectrum of crystals showed a strong broadening and shift to the red compared to the spectrum of the solubilized trimer (Fig. 3a, b); it also showed several distinct fluorescence peaks, indicating the presence of different emitting species arising from changes in chlorophyll configuration. This spectrum is similar to that found for LHCII oligomers¹⁷, the higher

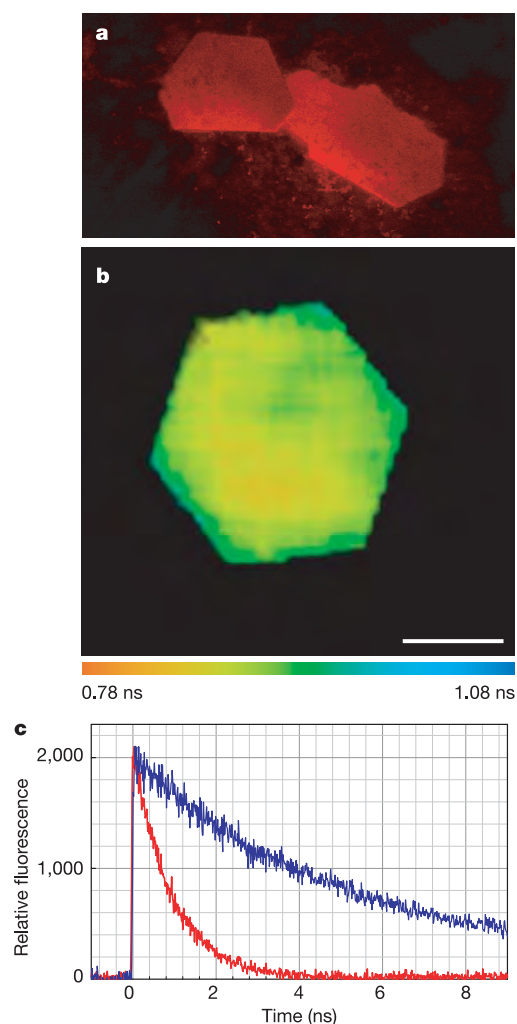


Figure 2 | Quenching of chlorophyll fluorescence in LHCII. **a**, Confocal fluorescence image of LHCII crystals. **b**, Fluorescence lifetime image of an LHCII crystal. The false colours indicate the fluorescence lifetimes in the corresponding pixels, and the colour code is presented below the panel. One picture consists of 64×64 pixels. Scale bar, $50 \mu\text{m}$. **c**, Fluorescence decay curves of trimers (blue trace) and crystals (red trace), the latter obtained by averaging decay curves from an area of 11×11 pixels within the crystal.

level of structure found in the spectrum of the crystal reflecting greater homogeneity of the crystal form.

Resonance Raman spectroscopy provided further evidence of major differences in the pigment configuration and pigment–protein interactions in the crystal compared to the trimer. In the spectra produced with 488.0 nm excitation (Fig. 3c), there was a significant enhancement of the modes at 951 and 955 cm^{-1} in the crystals, indicative of a twisting of the neoxanthin molecule as compared to the free trimer¹⁸. In LHCII crystals these modes are even more

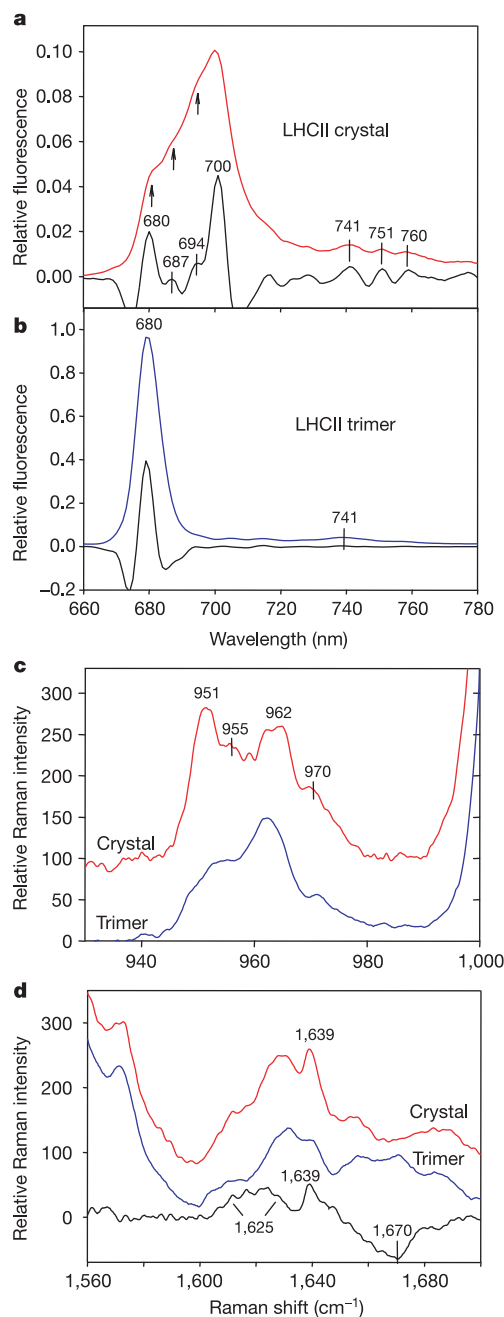


Figure 3 | Spectroscopic analysis of LHCII crystals. **a**, **b**, Low-temperature (77 K) fluorescence spectra of LHCII in crystalline (**a**, red) and trimeric (**b**, blue) forms, and second derivatives (black, multiplied by -1). Arrows show distinct shoulders on the crystal spectrum. **c**, **d**, Molecular configuration of LHCII-bound neoxanthin and chlorophyll *b* measured by resonance Raman spectroscopy of crystals (red) and trimers (blue). **c**, Excitation at 488.0 nm: neoxanthin-selective spectra. **d**, Excitation at 441.6 nm: chlorophyll *b* spectra. Also shown is the calculated difference spectrum (black, crystals minus trimers).

prominent than those previously found in oligomers^{18,19}. Similarly, the interactions between the chlorophyll *b* molecules and their environment are very different in the crystal compared to the trimer (Fig. 3d; note in particular the appearance of the prominent mode at $1,639\text{ cm}^{-1}$). The spectral changes show that the formyl carbonyl group of at least one chlorophyll *b*, free from interactions and in a non-polar environment in the trimer, becomes involved in a hydrogen bond in the crystal¹⁸. It is of interest that the mode at $1,639\text{ cm}^{-1}$ has an unusually small bandwidth ($6\text{--}7\text{ cm}^{-1}$), indicating a high degree of homogeneity in H-bonding strength for this new interaction. These changes in pigment configuration are observed in the near absence of protein–protein contacts and therefore can not be the result of inter-trimer interactions involving pigment molecules bound on the outside of the protein. For instance, the neoxanthin molecule only comes within 17 Å of the neighbouring trimer—the molecular twist cannot arise from any direct interaction (Fig. 1). We conclude that there is a protein conformational change leading to a reorganization of the structure within each of the interacting trimers.

Similarly, neither quenching nor the changes in the fluorescence spectrum can arise from the formation of inter-trimer associations of chlorophyll. The only pair of chlorophylls at close distance between trimers is chlorophyll *a* 614 and chlorophyll *b* 605 (Chl*a* 614 and Chl*b* 605; Fig. 1). The centre-to-centre distance between them is 12.25 Å , and the closest distance is 3.93 Å . It can be calculated that they are only weakly coupled⁷. Therefore, the only likely scenario is that a conformational change in LHCII upon crystallization causes subtle changes in the distances/orientations between its pigments, leading to the formation of quenching sites. The red-shift in fluorescence emission for LHCII crystals (Fig. 3a) suggests that these quenching centres could be chlorophyll dimers or excimers, which are known to have the potential to be powerful quenchers²⁰. There are a number of such pairs of chlorophylls observed in the LHCII structure. Figure 4 depicts two chlorophyll–chlorophyll pairs of particular significance. The first, on the stromal side of the complex, has been shown to be the site of lowest energy²¹ and can be referred to as the terminal emitter locus, which is an acceptor of the energy delivered to all LHCII pigments. It contains the Chl*a* 611/Chl*a* 612 pair, lutein 620 (Lut 620) and Chl*a* 610 (Fig. 4a). The presence of lutein in this domain is also consistent with the proposed role of carotenoids in energy dissipation⁶. The terminal emitter locus has previously been considered as a possible quenching site²².

A second chlorophyll pair, Chl*b* 606 and Chl*b* 607, is located on the luminal side of the complex and constitutes the most closely associated dimer (Fig. 4b). The closest distance between these molecules is 3.5 Å , between atoms of their macrocycles. The ligand of the central Mg of Chl*b* 607 is a water molecule (water 308), which hydrogen-bonds the formyl group of Chl*b* 606. In addition, the formyl group of Chl*b* 607 interacts with Gln 131, a residue that also forms an H-bond to the coordinating water 310 of Chl*b* 606 (ref. 7). These two chlorophylls thus constitute a special pair, with sandwiched interactions. Interestingly, these molecules are close to neoxanthin, the carotenoid whose configuration changes when LHCII adopts the dissipative state and which has been shown to have strong electronic interactions with the chlorophyll *b* molecules²³. The change observed in chlorophyll *b* interactions (Fig. 3d) could well be interpreted as the formation of the H-bond between water 308 and the formyl group of Chl*b* 606. Indeed, the unusual homogeneity of this H-bond is fully consistent with its playing an integral part in formation of the quenching centre. That this pair of chlorophyll molecules is the site of excitation quenching in LHCII is thus an attractive hypothesis.

The formation of the quenched antenna state, both in LHCII oligomers² and in NPQ *in vivo*²⁴, is controlled by the carotenoids of the xanthophyll cycle, violaxanthin and zeaxanthin. It should be noted that the crystal does not contain any zeaxanthin, which has been frequently proposed to be directly involved in NPQ. However, it is important to point out that the violaxanthin binding site is in close

proximity to the above two hypothetical quenching centres (Fig. 4c). Here violaxanthin (and also a phospholipid) is sandwiched between two chlorophylls, Chl*a* 611 and Chl*b* 601. Therefore, we now have a framework for understanding the roles of the various molecular partners that control NPQ. Exactly as predicted previously^{2,16}, de-epoxidation of violaxanthin into zeaxanthin, which stimulates energy dissipation, could modulate the structural changes in LHCII: the replacement of violaxanthin by zeaxanthin at this binding site could make the transition easier or pigment interactions could be further enhanced. The protein PsbS is necessary for the appearance of a part of NPQ *in vivo*²⁵, and it appears to be responsible for sensing the increase in acidification of the thylakoid lumen (ref. 26), the principal factor that signals the occurrence of excess light energy and therefore the need for transition to the dissipative state. We propose that at low lumen pH, PsbS interacts with LHCII in the photosynthetic membrane, promoting the conformational change and quenching in the same way as occurs *in vitro* upon oligomerization and crystallization.

In conclusion, the LHCII molecule behaves as a natural nano-switch that controls the emission or transfer of incoming light quanta. It can exist in very different functional states, the interconversion of which involve changes in pigment configuration brought about by a protein conformational change. We have provided the first

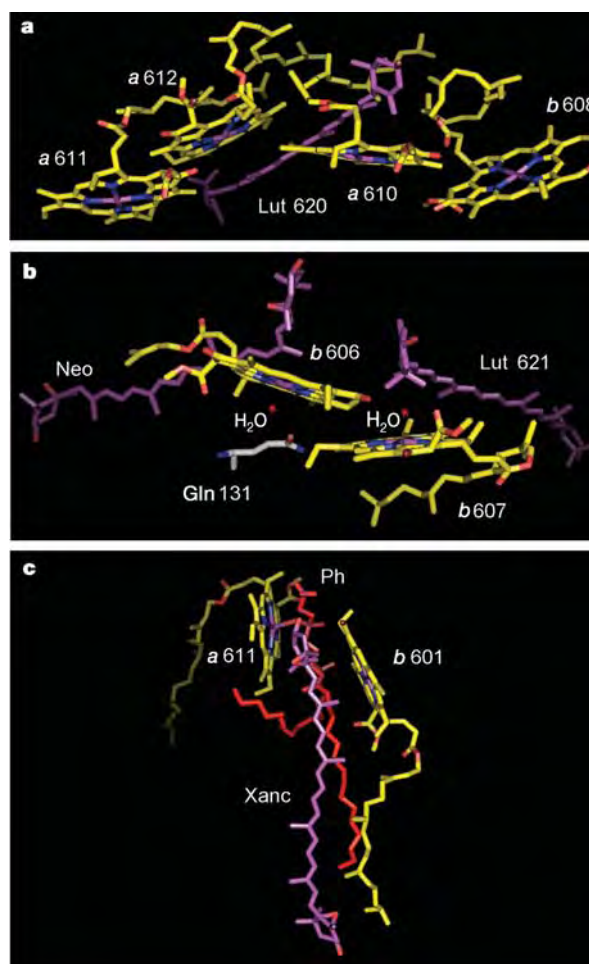


Figure 4 | Pigment-pigment interaction domains in LHCII. **a**, The terminal emitter domain; **b**, neoxanthin/chlorophyll *b* domain; **c**, the xanthophyll cycle carotenoid binding site. Carotenoids are shown in purple, phospholipid (Ph) in red, chlorophylls in yellow and a glutamine residue (Gln 131) in grey. Also shown are two water molecules (red spheres). Xanc, xanthophyll cycle carotenoid; Neo, neoxanthin; Lut, lutein. Chlorophylls (*a* 611, *b* 608, and so on) are named as in ref. 7

insights into the molecular design of a dissipative state of LHCII, probably made up of one or two chlorophyll pairs, which provides a channel for the safe dissipation of energy. Many questions are raised by these findings: how much of *in vivo* NPQ is due to this mechanism; what changes in protein conformation induce these pigment configurational changes; how is the extent of dissipation quantitatively controlled; are there other conformational states with different quenching strengths; does replacement of violaxanthin by zeaxanthin induce further structural change or just catalyse the conformational change; and what is the pigment configuration in the unquenched state?

METHODS

Crystals were obtained and soaked in a cryoprotectant solution as described⁷. Trimeric and oligomeric LHCII were obtained as described²⁷. Raman and fluorescence measurements were obtained from samples frozen on glass plates, as described^{17,28}. For crystalline LHCII, about 10 crystals were closely positioned in the centre of the excitation beam. Fluorescence emission spectra were measured using a SPEX Fluorolog FL3-22 spectrophotometer (Jobin-Yvon) equipped with xenon lamp excitation at 435 nm, defined by the double grating monochromator and photomultiplier detection. Resonance Raman spectra were measured using a Jobin-Yvon U1000 Raman spectrophotometer equipped with a liquid nitrogen-cooled CCD detector (Spectrum One, Jobin-Yvon). Excitation at 488.0 nm and 441.6 nm was provided by Coherent argon (Innova 100) and Liconix helium-cadmium lasers, respectively. Confocal images were taken with a Nikon TE300 inverted microscope. Excitation was with the 476 nm argon laser line, focused by a Plan Apochromat ×20 objective lens (numerical aperture 0.75) and fluorescence was measured with an internal detector through a HQ590LP long-pass filter, yielding images of 512 × 512 pixels (0.554 μm × 0.554 μm pixel size). For FLIM, a Bio-Rad Radiance 2100 MP system with a Nikon TE300 inverted microscope was used. A Ti:sapphire laser (Coherent Mira) pumped by a 5 W Coherent Verdi laser generated two-photon excitation pulses (860 nm, 150 fs) at a repetition rate of 76 MHz. The excitation light (0.3 mW) was directly coupled into the microscope and focused into the sample using a Plan Apochromat ×20 objective lens (numerical aperture 0.75). Fluorescence light was detected using non-descanned single photon counting detection with a Hamamatsu R3809U MCP PMT, with a time resolution of ~50 ps (in a time window of 12.5 ns, with 1,024 channels), through a HQ700/75m bandpass filter, yielding images of 62 × 62 pixels (3.5 μm × 3.5 μm). The fluorescence decay curves of each pixel were fitted with a triple exponential decay model. The lifetimes did not change with excitation densities up to 16 times higher, excluding singlet–singlet and singlet–triplet annihilation, which may lead to decreased lifetimes²⁹.

Received 14 April; accepted 9 May 2005.

- Demmig-Adams, B. & Adams, W. W. III. Antioxidants in photosynthesis and human nutrition. *Science* **298**, 2149–2153 (2002).
- Horton, P., Ruban, A. V. & Walters, R. G. Regulation of light harvesting in green plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47**, 655–684 (1996).
- Niyogi, K. K. Photoprotection revisited. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **50**, 333–359 (1999).
- Külheim, C., Ågren, J. & Jansson, S. Rapid regulation of light harvesting and plant fitness in the field. *Science* **297**, 91–93 (2002).
- Barber, J. & Andersson, B. Too much of a good thing: light can be bad for photosynthesis. *Trends Biochem. Sci.* **17**, 61–66 (1992).
- van Amerongen, H. & van Grondelle, R. Understanding the energy transfer function of LHCII, the major light-harvesting complex of green plants. *J. Phys. Chem. B* **105**, 604–617 (2001).
- Liu, Z. *et al.* Crystal structure of spinach major light-harvesting complex at 2.72 Å resolution. *Nature* **428**, 287–292 (2004).
- Muller, P., Li, X. P. & Niyogi, K. K. Non-photochemical quenching. A response to excess light energy. *Plant Physiol.* **125**, 1558–1566 (2001).
- Horton, P. & Ruban, A. V. Molecular design of the photosystem II light harvesting antenna: photosynthesis and photoprotection. *J. Exp. Bot.* **56**, 365–373 (2005).
- Holt, N. E. *et al.* Carotenoid cation formation and the regulation of photosynthetic light harvesting. *Science* **307**, 433–436 (2005).
- Finazzi, G. *et al.* A zeaxanthin-independent nonphotochemical quenching mechanism localised in the photosystem II core complex. *Proc. Natl Acad. Sci. USA* **101**, 12375–12380 (2004).
- Standfuss, J. *et al.* Mechanisms of photoprotection and nonphotochemical quenching in pea light-harvesting complex at 2.5 Å resolution. *EMBO J.* **24**, 919–928 (2005).
- Moya, I. *et al.* Time-resolved fluorescence analysis of the photosystem II antenna proteins in detergent micelles and liposomes. *Biochem* **40**, 12552–12561 (2001).
- Mullet, J. E. & Arntzen, C. J. Simulation of grana stacking in a model membrane system. *Biochim. Biophys. Acta* **589**, 100–117 (1980).
- Mullineaux, C. W. *et al.* Excitation-energy quenching in aggregates of the LHCII chlorophyll-protein complex: a time-resolved fluorescence study. *Biochim. Biophys. Acta* **1141**, 23–28 (1993).
- Horton, P. *et al.* Control of the light harvesting function of chloroplast membranes by aggregation of the LHCII chlorophyll protein complex. *FEBS Lett.* **292**, 1–4 (1991).
- Ruban, A. V. & Horton, P. Mechanism of ΔpH-dependent dissipation of absorbed excitation energy by photosynthetic membranes. I. Spectroscopic analysis of isolated light harvesting complexes. *Biochim. Biophys. Acta* **1102**, 30–38 (1992).
- Ruban, A. V., Horton, P. & Robert, B. Resonance Raman spectroscopy of the photosystem II light harvesting complex of green plants. A comparison of the trimeric and aggregated states. *Biochemistry* **34**, 2333–2337 (1995).
- Robert, B., Horton, P., Pascal, A. A. & Ruban, A. V. Insights into the molecular dynamics of the plant light harvesting proteins *in vivo*. *Trends Plant Sci.* **9**, 385–390 (2004).
- Beddard, G. S. & Porter, G. Concentration quenching in chlorophyll. *Nature* **260**, 366–367 (1976).
- Remelli, R. *et al.* Chlorophyll binding to monomeric light-harvesting complex. *J. Biol. Chem.* **274**, 33510–33521 (1999).
- Wentworth, M., Ruban, A. V. & Horton, P. Thermodynamic investigation into the mechanism of the chlorophyll fluorescence quenching in isolated photosystem II light harvesting complexes. *J. Biol. Chem.* **278**, 21845–21850 (2003).
- Croce, R. *et al.* The neoxanthin binding site of the major light harvesting complex (LHCII) from higher plants. *FEBS Lett.* **456**, 1–6 (1999).
- Demmig-Adams, B. Carotenoids and photoprotection in plants: a role for the xanthophyll zeaxanthin. *Biochim. Biophys. Acta* **1020**, 1–24 (1990).
- Li, X. P. *et al.* A pigment-binding protein essential for regulation of photosynthetic light harvesting. *Nature* **403**, 391–395 (2000).
- Li, X. P. *et al.* Regulation of photosynthetic light harvesting involves intrathylakoid lumen pH sensing by the PsbS protein. *J. Biol. Chem.* **279**, 22866–22874 (2004).
- Ruban, A. V. *et al.* Determination of the stoichiometry and strength of binding of xanthophylls to the photosystem II light harvesting complexes. *J. Biol. Chem.* **274**, 10458–10465 (1999).
- Ruban, A. V., Pascal, A. A. & Robert, B. Xanthophylls of the major photosynthetic light-harvesting complex of plants: identification, conformation and dynamics. *FEBS Lett.* **477**, 181–185 (2000).
- Barzda, V. *et al.* Fluorescence lifetime heterogeneity in aggregates of LHCII revealed by time-resolved microscopy. *Biophys. J.* **81**, 538–546 (2001).

Acknowledgements This work was supported by UK Biotechnology and Biological Sciences Research Council, the INTRO2 European Union FP6 Marie Curie Research Training Network, The Royal Society, Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) through the research programme of the Stichting voor Fundamenteel Onderzoek der Materie (FOM), President Foundation of CAS, the Knowledge Innovation Project of CAS, and the National Natural Sciences Foundation of China. The authors thank J. W. Borst of the MicroSpectroscopy Centre at Wageningen University for help in enabling and performing the FLIM measurements and for discussions.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.H. (p.horton@sheffield.ac.uk) or B.R. (robert@dsvidf.cea.fr).

LETTERS

X-ray structure of a tetranucleosome and its implications for the chromatin fibre

Thomas Schalch¹, Sylwia Duda¹, David F. Sargent¹ & Timothy J. Richmond¹

DNA in eukaryotic chromosomes is organized in arrays of nucleosomes compacted into chromatin fibres. This higher-order structure of nucleosomes is the substrate for DNA replication, recombination, transcription and repair. Although the structure of the nucleosome core is known at near-atomic resolution¹, even the most fundamental information about the organization of nucleosomes in the fibre is controversial. Here we report the crystal structure of an oligonucleosome (a compact tetranucleosome) at 9 Å resolution, solved by molecular replacement using the nucleosome core structure. The structure shows that linker DNA zigzags back and forth between two stacks of nucleosome cores, which form a truncated two-start helix, and does not follow a path compatible with a one-start solenoidal helix². The length of linker DNA is most probably buffered by stretching of the DNA contained in the nucleosome cores. We have built continuous fibre models by successively stacking tetranucleosomes one on another. The resulting models are nearly fully compacted and most closely resemble the previously described crossed-linker model³. They suggest that the interfaces between nucleosomes along a single helix start are polymorphic.

The nucleosome is the fundamental repeating element of chromatin and comprises between 157 and 240 base pairs (bp) of DNA, the four core histone proteins (H2A, H2B, H3 and H4) and the linker histone protein (H1). The nucleosome core contains 147 bp of DNA supercoiled in 1.67 left-handed turns around a core histone octamer⁴. When native chromatin is examined by electron microscopy under conditions of low ionic strength, nucleosomes appear as 11-nm beads on a string^{5,6}. On addition of salt or when observed *in situ*⁷, nucleosome arrays appear as compact fibres approximately 30 nm in diameter. Subsequent to electron microscopy and small-angle X-ray scattering (SAXS) studies, two classes of models for the chromatin fibre emerged: the one-start solenoidal helix, in which a linear array of nucleosomes is coiled², and the two-start helix, in which nucleosomes are assembled in a zigzag ribbon that twists or supercoils^{3,8}. Evidence cited in favour of the solenoid model is its invariant diameter with respect to DNA linker length^{9,10}, increased compaction for fibres with six or more nucleosomes¹¹, and a possible requirement for supercoiled linker DNA¹². Evidence cited in support of two-start helix models is, conversely, variation of fibre diameter with DNA linker length¹³, a zigzag path of nucleosomes as seen from tomographic reconstruction *in situ*¹⁴, and linker DNA that is not tightly bent¹⁵.

We have determined, at low resolution, the first structure of an oligonucleosome using a tetranucleosome assembled *in vitro*. Chromatin reaches maximum compaction *in vitro* in the presence of multivalent cations or histone H1 (for example, see refs 16, 17), and therefore crystals were grown from solutions having a divalent cation concentration yielding compact chromatin. The tetranucleosome comprises two stacks of two nucleosome cores (N1 and N2, and N1' and N2') with three segments of linker DNA (LB, LB' and LS)

connecting them (Fig. 1a, b). A crystallographic two-fold axis passes between the stacks making N1, N2 and LB identical to N1', N2' and LB'. The two-fold axis bisects LS. The two stacks are 146.1 Å apart, centre to centre, measured as the distance between the two stack axes passing through the centres of N1 and N2 and of N1' and N2' (the nucleosome core centre is the intersection of its dyad and superhelix axes⁴) (Fig. 1c). Notably, the stacks are rotated by -71.3° (left-handed twist) with respect to each other around the shortest line segment connecting their stack axes. The N1 and N2 nucleosome cores within the same stack are separated by 57.6 Å centre to centre, and are oriented with the centre of their DNA (dyad position) approximately facing the opposite stack. Their superhelix axes and dyad axes are at angles of 14.9° and -21.7° , respectively. N1 and N2 interact via their octamer surfaces without direct connection via linker DNA. Overall, the tetranucleosome has dimensions of approximately $120 \times 150 \times 250$ Å and corresponds to a two-start helix, consistent with our recent crosslinking and electron microscopy study¹⁷.

The tetranucleosome used here incorporates DNA having four 147-bp copies of the '601' sequence, which yields a single high-affinity position for the histone octamer (Supplementary Methods). The nucleosome core-forming segments are connected by 20-bp DNA linkers. The 167-bp repeat length produced is approximately equal to the chromatin repeat found for *Saccharomyces cerevisiae*¹⁸ and for neurons in the cerebral cortex¹⁹. Tetranucleosomes were assembled using recombinant *Xenopus laevis* histone octamers lacking post-translational modifications²⁰. On addition of divalent cations, these tetranucleosomes compact into homogeneous particles as judged by sedimentation analysis (T.S. and T.J.R., unpublished data), as also occurs for nucleosome arrays with more and longer repeats¹⁶. Crystals were obtained under conditions providing maximum array compaction¹⁶. They are stable between 20–60 mM MgCl₂ and withstand addition of up to 150 mM KCl. The structure was determined under conditions approximating physiological ionic strength.

The tetranucleosome structure was solved by molecular replacement using the nucleosome core particle structure¹ (Supplementary Table 1). A single stack of two nucleosomes constitutes the crystal asymmetric unit. However, each asymmetric unit is overlaid with a second copy, rotated 180° , according to the *I*222 space group symmetry. This two-fold disorder results in the superposition of two tetranucleosomes (Supplementary Fig. 1). Importantly, the electron density for the linker DNA segments connecting stacks is sufficiently clear for both orientations to permit reliable interpretation (Fig. 2a). The independent positional parameters of the two nucleosomes in the asymmetric unit were refined simultaneously to reveal the location of LB, LB' and LS (Fig. 2b).

A consequence of the molecular two-fold symmetry of the tetranucleosome is that there are only two types of linker DNA segments: LS (straight) and LB (bent) (Fig. 1a, b). LB is maximally bent by 35°

¹ETH Zürich, Institute for Molecular Biology and Biophysics, ETH-Hönggerberg, CH-8093 Zürich, Switzerland.

in essentially one direction, inclusive of 6 bp from each adjoining nucleosome core. The bending occurs into the minor groove primarily at the T·A base pair step at the centre of LB. LS and LB are not involved in crystal contacts. The presence of two different linker conformations suggests that a dinucleosome may best represent the repeating unit of the higher-order structure. Both linkers were modelled using a 20-bp B-form DNA segment corresponding to the linker length contained in the tetranucleosome DNA repeat. The entire tetranucleosome DNA was regularized by energy minimization to smooth the transitions between the linker and nucleosome core DNA. Using 31-bp steps (20 bp from the linker and 6 bp from each adjoining nucleosome core), the mean rise and twist per base pair are 3.54 Å (LS) and 3.48 Å (LB), and 10.20 bp per turn (LS) and 10.14 bp per turn (LB), respectively. Calculations using 32-bp steps instead yield 3.38 Å (LB) and 3.43 Å (LS), and 10.47 bp per turn (LB) and 10.53 bp per turn (LS), values matching well those accepted for free B-form DNA, and suggesting that 1 bp per linker has been pulled out of the nucleosome cores. A mechanism for buffering linker DNA

length and twist using nucleosome core DNA has been described⁴.

We have used the tetranucleosome structure, which does not form continuous fibres in the crystals, to build 'direct' and 'idealized' models of the chromatin fibre. The fibre axis for both models is chosen to be orthogonally intersecting the tetranucleosome two-fold axis and bisecting the angle formed by the two stack axes (Fig. 1c). In the structure, N1 and N2 are related by a -38.1° rotation around the fibre axis and a 33.4 Å translation along it. Therefore, the direct model with a repeating unit of N1–N2 was constructed by placing successive tetranucleosomes along the fibre axis using twice these values (Fig. 3a). The maximum interfacial angle resulting between nucleosomes in adjacent tetranucleosomes is 39.3° compared to 14.9° between N1 and N2 within the tetranucleosome, and as a consequence, unacceptable steric interference occurs between the linker and nucleosome core DNA of neighbouring tetranucleosomes. Notwithstanding the steric overlap in the direct model, it contains 18.9 nucleosomes per turn in a period of 316.0 Å, yielding 6.6 nucleosomes per 110 Å (nucleosome core diameter), consistent with measured values for mass per unit length of the chromatin fibre²¹. The fibre diameter is 24–25 nm. Steric interference is most simply relieved by increasing the separation of tetranucleosomes, which decreases the number of nucleosomes per 110 Å to 5.8.

We built the idealized model to eliminate the steric overlap of the direct model while maintaining its helix parameters. It is constructed with a single nucleosome as the repeating unit, thereby equalizing all nucleosome interfacial angles. N2 was first rotated by 9° to make its dyad axis intersect the fibre axis orthogonally. The same repeat

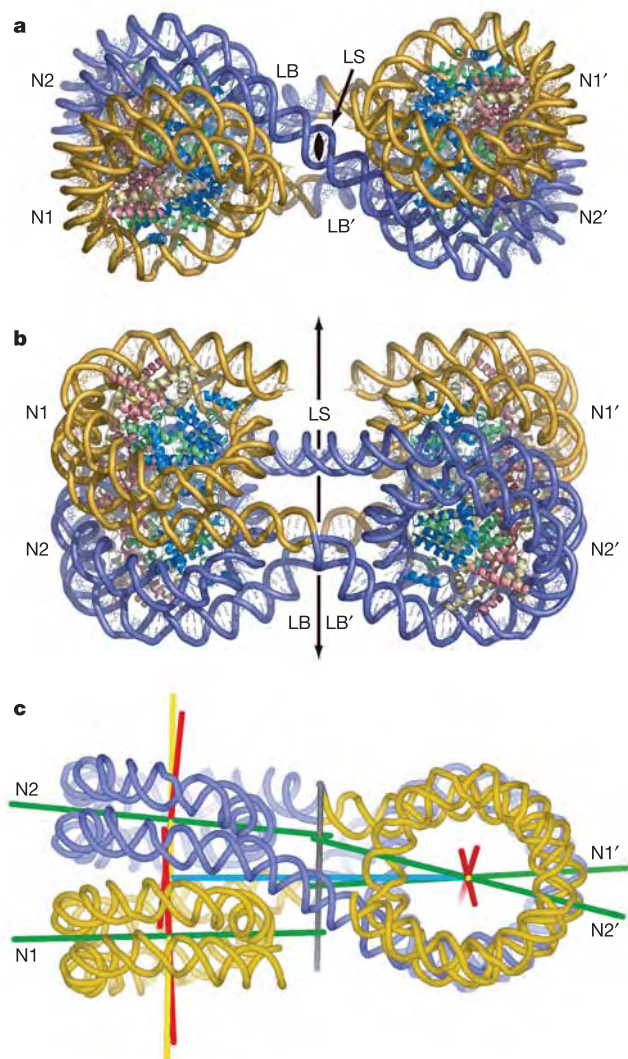


Figure 1 | Tetranucleosome structure. **a**, View down a crystallographic two-fold axis (black oval) relating the nucleosomes N1 and N2 to N1' and N2' and bent linker DNA segment LB to LB'. The two-fold axis passes through the straight linker DNA segment LS (arrow). **b**, View orthogonal to the two-fold axis in **a** by rotation about the tetranucleosome long axis. **c**, Definition of axes: nucleosome core dyad (green) and superhelix axes (red); tetranucleosome stack axes (yellow); line segment connecting stack axes (cyan); fibre axis (grey). (See Supplementary Video 1.)

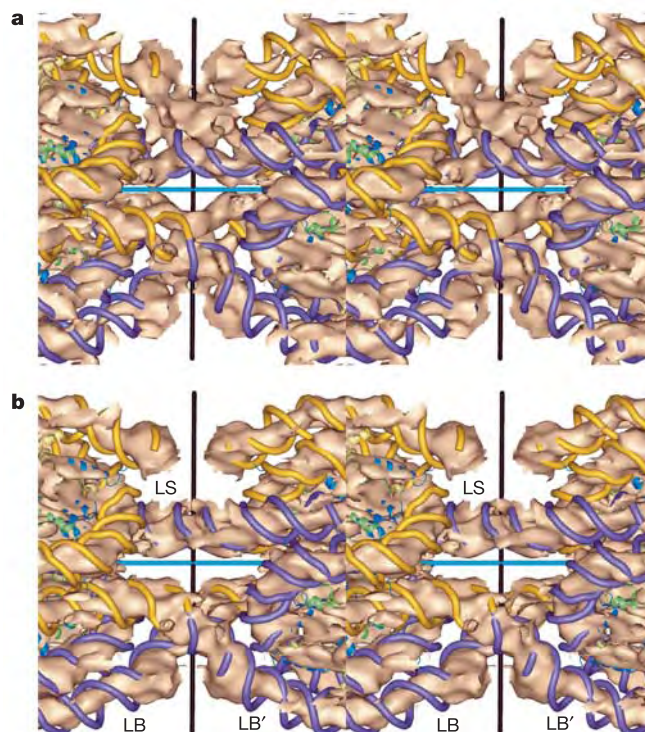


Figure 2 | Electron density for the linker DNA (stereographs). View direction and labels are as for Fig. 1b. The $2F_o - F_c$ electron-density surface is set at 0.7σ . Traces for the DNA chains for one tetranucleosome are shown (gold, blue). **a**, Linker DNA omit map calculated using phases from the nucleosome core particle (10 bp from both DNA termini were omitted). The surface corresponds to two superimposed copies of the tetranucleosome electron density related by a crystallographic two-fold axis (cyan). The surface is cut off if more distant than 8 Å from both tetranucleosome models. **b**, Map calculated using phases from the complete model. The surface is cut off if more distant than 8 Å from the single tetranucleosome model shown.

parameters as for the observed N1–N2 step were applied (Fig. 3b). Limited steric overlap occurs, and is relieved by increasing the rotation angle of N2 around its dyad axis from the observed 39.9° to 70° (dihedral angle between the nucleosome superhelix and fibre axes; 0° corresponds to parallel axes). This larger angle is consistent with the range of values ($\gamma_n = 57\text{--}77^\circ$) determined from measurements of the photochemical dichroism for several native chromatin samples, each having a different linker DNA length²². DNA connectivity within the idealized model requires limited flexibility within the linker DNA and up to 10 bp of terminal nucleosome core DNA. The resulting angle between the segment of straight linker DNA and the fibre axis matches the value ($\beta_1 = 60^\circ \pm 3^\circ$) from the dichroism study for HeLa chromatin ($\sim 20\text{-bp}$ linker)²². The models can explain the appearance of the transition in sedimentation behaviour of short chromatin oligomers occurring at a length of about six nucleosomes¹¹. Nucleosomes in the two-helix starts coming closest together are related i to $i + 5$ and $i + 7$ (contact repeat) and could make stabilizing interactions. SAXS distributions calculated from the models display peaks at 5.6 and 12.5 nm compared to the characteristic peaks of native chromatin at 5.6 and 11 nm, explained by the nucleosome repeat distance along the helix path and by the spacing between helix gyres, respectively²³ (Supplementary Fig. 2). The larger value for the gyre spacing of the models suggests that the two-start helix could be further twisted to bring nucleosomes in neighbouring gyres into contact.

The two-start helices described by the direct and idealized models are twisted ribbons that follow a straight fibre axis. The left-handed twist is a feature of the tetranucleosome structure. If, instead of being straight, the fibre axis were to take up a left-handed supercoiled path with each nucleosome maintaining its orientation with respect to the

axis, the two stacks of nucleosomes would unwind as dictated by the linking number equation. One turn of the intertwined stacks would be converted to one left-handed supercoil. The purely twisted form of the fibre shown in our models corresponds to the crossed-linker model³, and the purely supercoiled form corresponds to the helical ribbon model⁸. Intermediate geometries having contributions from fibre twisting and supercoiling are possible. Assuming a constant spacing of nucleosomes along the fibre axis, the chromatin fibre would probably increase its radius and shorten its repeat period on going from the twisted to the supercoiled form (for example, as for the conversion of B-form to A-form DNA). The radial location of the linker histone H1 measured by neutron scattering from reconstituted chromatin fibres favours the twisted over the superhelical form of the fibre for the bulk of chromatin²⁴.

The N1–N2 interface within the tetranucleosome structure displays relatively strong density where adjacent H2A–H2B dimers meet (Fig. 3c). Elements of the N1 and N2 H2A– $\alpha 2$, H2B– $\alpha 1$ and H2B– αC helix elements make a pseudo-two-fold symmetric interaction in which both H2A–H2B dimers could be shifted out of the nucleosome by up to $5\text{ \AA}$. This interface does not permit the inter-nucleosomal interaction between the base of the H4 tail (amino acids 16–26) and H2A–H2B dimer seen in the nucleosome core particle structure¹ and occurring in compacted nucleosome arrays^{16,17}. In contrast to the N1–N2 and N2–N1 interfaces in the direct fibre model, the interface in the idealized model places the H4 tail and H2A–H2B dimer regions in sufficiently close proximity to allow interaction (Fig. 3d). These models suggest that packing of nucleosomes in the chromatin fibre is polymorphic. Each different arrangement could affect the conformation of linker DNA.

Questions of how longer or non-uniform linker DNA lengths are

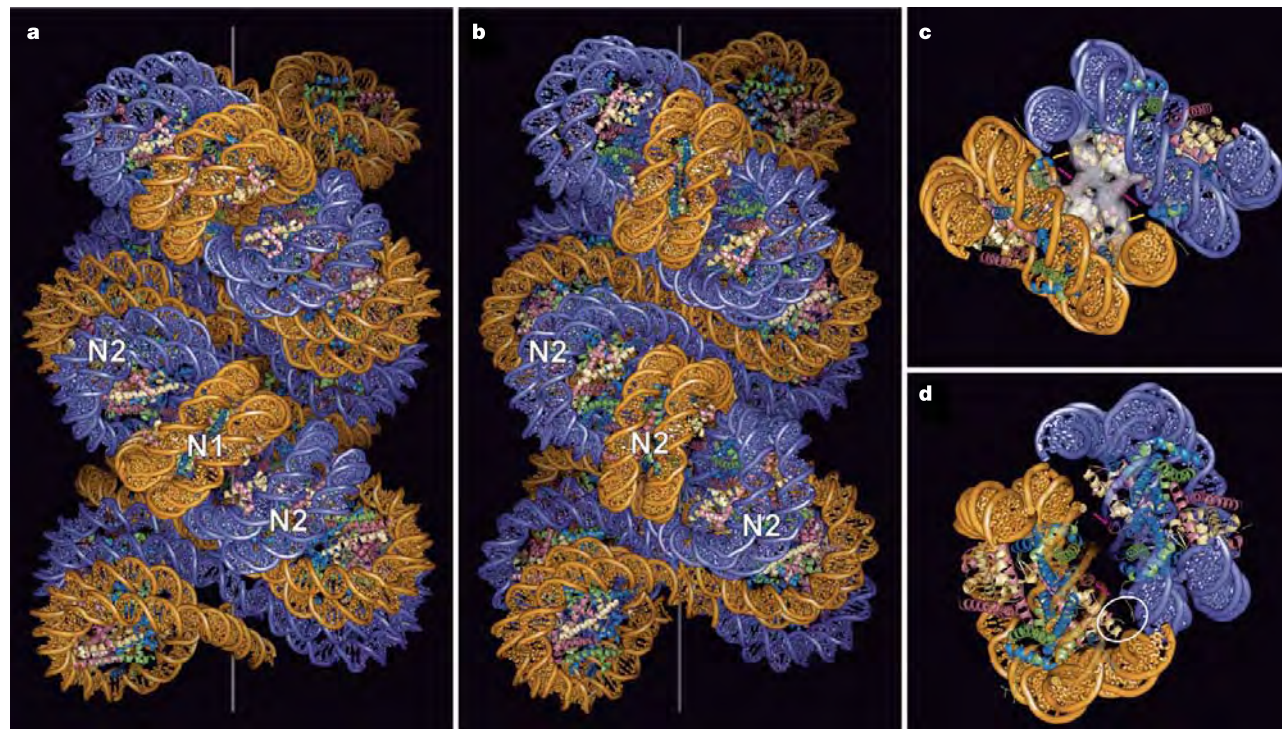


Figure 3 | Chromatin fibre models. **a**, Direct model built using the rotation and translation parameters relating nucleosome N1 (gold) to N2 (blue) to generate the tetranucleosome interface N2 to N1. The fibre axis (grey) is chosen to make both helix starts equivalent. **b**, Idealized model built using N2 (blue and gold) oriented to have its dyad axis orthogonally intersect the fibre axis. All nucleosome interfaces are thereby equivalent. The rotation of N2 is increased to avoid steric overlaps. **c**, The interface between N1 (gold) and N2 (blue) in the direct model (**a**) viewed from inside the fibre. Electron density for the interacting regions of the H2A–H2B dimers is shown as for

Fig. 2. The H2A– $\alpha 2$ helices (yellow arrows) and H2B– αC helices (red arrows) form part of a pseudo-two-fold symmetric interaction. **d**, The interface between nucleosomes, which are all equivalent, in the idealized model (**b**) viewed from inside the fibre. Increasing the rotation of the nucleosome around its dyad axis brings into proximity the regions (circled) of the H4 tail and H2A–H2B dimer surface crosslinked in a previous study¹⁷. The positions of the H2B– αC helices (red arrows) are indicated for comparison to the direct model (**c**).

accommodated in twisted two-start fibres are not conclusively answered by the models presented. As proposed previously, longer lengths of DNA crossing back and forth between the two helix starts could be included simply by increasing fibre diameter¹³. However, this also has the consequence of increasing the contact repeat substantially. As seen from our further model building, both the diameter and contact repeat variation with linker length are reduced because linker trajectories are not orthogonal to the fibre axis. Nevertheless, continuance of a constant radius and contact repeat along the chromatin fibre would require significant uniformity of linker length. Variations of up to ± 5 bp could be absorbed locally through adjustment of DNA length on adjacent nucleosome cores. As judged from analyses of chromatin assembled *in vitro* using cell extracts, this degree of uniformity may be prevalent in native bulk chromatin^{25,26}. Variations of roughly ± 10 bp may be accommodated by polymorphic packing within each nucleosome stack. Larger, abrupt changes, perhaps arising from relatively strong positioning sequences for two successive nucleosomes, would cause a structural perturbation in the fibre. Such a deviation might promote nuclear factor access to a nucleosome in one stack while the other stack maintains the overall form of the fibre. Conversely, substantial linker length deviations may facilitate higher-level folding in an otherwise highly compact fibre. Further structural studies using highly defined chromatin samples will enable these questions to be answered.

METHODS

Sample preparation and crystallization. Tetranucleosomes were prepared as described for longer arrays¹⁶, and stored as a 6 mg ml^{-1} solution in 10 mM KCl, 0.25 mM EDTA and 10 mM Tris-Cl, pH 7.5. Crystals were grown by vapour diffusion, mixing the storage solution with an equal volume of 50 mM KCl, 180 mM MgCl_2 and 10 mM K-cacodylate, pH 6.0 solution, and equilibrating against the same solution $\times 4$ diluted. Crystals were transferred to 26% 2-methyl-2,4-pentanediol, 12.5 mM KCl, 40 mM MgCl_2 and 10 mM K-cacodylate, pH 6.0 before cryo-cooling in liquid propane (-120°C) and storing under liquid nitrogen.

Structure determination. X-ray data were collected using the X06SA beamline at the Swiss Light Source (Villigen PSI, Switzerland) (Supplementary Table 1). Molecular replacement solutions were found using Beast²⁷ and indicated a two-fold disorder in crystal packing (Supplementary Fig. 1). Rigid body refinement was done with Refmac5 (ref. 28). The 'simple' scaling algorithm (no bulk solvent correction) in Refmac5 was used to scale F_c to F_o . Unbiased electron density maps were calculated using σ_A -weighted structure factors output from Refmac5. The junctions of linker DNA and nucleosome core DNA were regularized using CNS²⁹. Model building was done with O³⁰. Illustrations were prepared with PyMOL (<http://www.pymol.org/>).

Received 10 February; accepted 22 April 2005.

- Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.* **319**, 1097–1113 (2002).
- Finch, J. T. & Klug, A. Solenoidal model for superstructure in chromatin. *Proc. Natl Acad. Sci. USA* **73**, 1897–1901 (1976).
- Williams, S. P. *et al.* Chromatin fibers are left-handed double helices with diameter and mass per unit length that depend on linker length. *Biophys. J.* **49**, 233–248 (1986).
- Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003).
- Oudet, P., Gross-Bellard, M. & Chambon, P. Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. *Cell* **4**, 281–300 (1975).
- Thoma, F. & Koller, T. Influence of histone H1 on chromatin structure. *Cell* **12**, 101–107 (1977).
- Ris, H. & Korenberg, J. in *Cell Biology* (eds Prescott, D. M. & Goldstein, L.) 268–361 (Academic, New York, 1979).
- Woodcock, C. L., Frado, L. L. & Rattner, J. B. The higher-order structure of chromatin: evidence for a helical ribbon arrangement. *J. Cell Biol.* **99**, 42–52 (1984).
- Widom, J., Finch, J. T. & Thomas, J. O. Higher-order structure of long repeat chromatin. *EMBO J.* **4**, 3189–3194 (1985).
- Woodcock, C. L. Chromatin fibers observed *in situ* in frozen hydrated sections. Native fiber diameter is not correlated with nucleosome repeat length. *J. Cell Biol.* **125**, 11–19 (1994).
- Butler, P. J. & Thomas, J. O. Changes in chromatin folding in solution. *J. Mol. Biol.* **140**, 505–529 (1980).
- McGhee, J. D., Nickol, J. M., Felsenfeld, G. & Rau, D. C. Higher order structure of chromatin: Orientation of nucleosomes within the 30 nm chromatin solenoid is independent of species and spacer length. *Cell* **33**, 831–841 (1983).
- Athey, B. D., Smith, M. F., Rankert, D. A., Williams, S. P. & Langmore, J. P. The diameters of frozen-hydrated chromatin fibers increase with DNA linker length: evidence in support of variable diameter models for chromatin. *J. Cell Biol.* **111**, 795–806 (1990).
- Horowitz, R. A., Agard, D. A., Sedat, J. W. & Woodcock, C. L. The three-dimensional architecture of chromatin *in situ*: electron tomography reveals fibers composed of a continuously variable zig-zag nucleosomal ribbon. *J. Cell Biol.* **125**, 1–10 (1994).
- Pehrson, J. R. Thymine dimer formation as a probe of the path of DNA in and between nucleosomes in intact chromatin. *Proc. Natl Acad. Sci. USA* **86**, 9149–9153 (1989).
- Dorigo, B., Schalch, T., Bystricky, K. & Richmond, T. J. Chromatin fiber folding: requirement for the histone H4 N-terminal tail. *J. Mol. Biol.* **327**, 85–96 (2003).
- Dorigo, B. *et al.* Nucleosome arrays reveal the two-start organization of the chromatin fiber. *Science* **306**, 1571–1573 (2004).
- Thomas, J. O. & Furber, V. Yeast chromatin structure. *FEBS Lett.* **66**, 274–280 (1976).
- Allan, J., Rau, D. C., Harborne, N. & Gould, H. Higher order structure in a short repeat length chromatin. *J. Cell Biol.* **98**, 1320–1327 (1984).
- Luger, K., Rechsteiner, T. J., Flaus, A. J., Wayne, M. M. & Richmond, T. J. Characterization of nucleosome core particles containing histone proteins made in bacteria. *J. Mol. Biol.* **272**, 301–311 (1997).
- Gerchman, S. E. & Ramakrishnan, V. Chromatin higher-order structure studied by neutron scattering and scanning transmission electron microscopy. *Proc. Natl Acad. Sci. USA* **84**, 7802–7806 (1987).
- Sen, D., Mitra, S. & Crothers, D. M. Higher order structure of chromatin: evidence from photochemically detected linear dichroism. *Biochemistry* **25**, 3441–3447 (1986).
- Widom, J. & Klug, A. Structure of the 300 Å chromatin filament: X-ray diffraction from oriented samples. *Cell* **43**, 207–213 (1985).
- Graziano, V., Gerchman, S. E., Schneider, D. K. & Ramakrishnan, V. Histone H1 is located in the interior of the chromatin 30-nm filament. *Nature* **368**, 351–354 (1994).
- Blank, T. A. & Becker, P. B. Electrostatic mechanism of nucleosome spacing. *J. Mol. Biol.* **252**, 305–313 (1995).
- Korber, P. & Horz, W. *In vitro* assembly of the characteristic chromatin organization at the yeast PHO5 promoter by a replication-independent extract system. *J. Biol. Chem.* **279**, 35113–35120 (2004).
- Read, R. Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallogr. D* **57**, 1373–1382 (2001).
- Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
- Brünger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Schulze-Bries, T. Tomizaki and A. Wagner for assistance in data collection, and J. Widom for providing us with the 601 sequence.

Author Information Coordinates have been deposited in the Protein Data Bank with identifier 1zbb. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.J.R. (richmond@mol.biol.ethz.ch).

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

They'll see you now

If you're like most people, waiting for a performance review is akin to sitting in reception at the dentist's. You hope your appointment isn't going to result in any pain, but you expect the worst.

Perhaps some scientists aspire to academia in order to avoid such meddlesome interactions, but two UK lab heads say that feedback is a valuable part of the scientific process. The trick, they say, is making these routine check-ups a two-way process. They advise keeping the meetings relatively informal and focused on progress, rather than deficiency.

Kay Davies is head of the University of Oxford's anatomy department, honorary director of the Medical Research Council's Functional Genetics Unit and co-director of the Oxford Centre for Gene Function. She makes appraisals an annual event for the scientists she directs, and says they keep her researchers focused on specific objectives. Yearly meetings help them to see what sort of research opportunities exist and how to pick up the necessary skills and resources, she adds.

A different strategy is used by Steve Jackson, who heads

a lab at the Gurdon Institute of Cancer and Developmental Biology in Cambridge. He conducts informal exit interviews when students and postdocs end their stint. These casual conversations give the young scientists career goals and provide Jackson with valuable feedback about his lab's direction. He has also used the chats to find employees for KuDOS Pharmaceuticals, a biotechnology company that he created in 1997.

Labs all over the world differ in their policy towards such interactions. Some provide few opportunities for appraisals or feedback, others rely on casual conversations, still others apply a more formal framework. But in science, as in dentistry, regular appointments — whether initiated by you or your supervisor — can improve your long-term health.



Paul Smaglik, *Naturejobs* editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

US Head Office, New York
 345 Park Avenue South, 10th Floor,
 New York, NY 10010-1707
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager/Corporations:
 Peter Bless
 Classified Sales Representatives
 Tel: +1 800 989 7718

**New York/Pennsylvania/
 Latin America:** Kelly Roman
**Midwest USA/Maryland/
 NIH:** Wade Tucker
East USA/Canada:
 Janine Taormina

San Francisco Office
Classified Sales Representative:
 Michaela Bjorkman
 225 Bush Street, Suite 1453
 San Francisco, CA 94104
 Tel: +1 415 781 3803
 Fax: +1 415 781 3805
 e-mail: m.bjorkman@naturesf.com

European Head Office, London
 The Macmillan Building,
 4 Crinan Street,
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs Sales Director: Nevin Bayoumi (4978)
European Sales Manager: Andy Douglas (4975)

Advertising Production Manager: Billie Franklin
 To send materials use London address above.
 Tel: +44 (0) 20 7843 4814
 Fax: +44 (0) 20 7843 4814
 e-mail: naturejobs@nature.com

Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

European Satellite Office
 Patrick Phelan
 e-mail: p.phelan@nature.com

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Golden opportunities

No longer rivals, Oxford, Cambridge and London are now working towards a common goal — ensuring the 'golden triangle' becomes a global science hub. **Paul Smaglik** investigates.

Rivalries. The universities in the southeast of England are rife with them. Cambridge and Oxford have competed for academic glory throughout their august histories. More recently, both have vied for bragging rights with their bigger, brasher competitor — London. And even within London, Imperial College, University College and King's College have long jockeyed for top position.

Now, out of necessity or perhaps a realization that the real competition lies outside Britain, these sometime rivals are combining forces in a number of ways — strategic alliances, joint appointments and shared infrastructure among them.

Ken Powell, who left a professorship at the Wolfson Institute for Biomedical Research at University College London to found Arrow Therapeutics in the city, says such cooperation is long overdue, and still in woefully short supply. "There needs to be more collaboration,"

The way ahead: Oxford, which now receives more than US\$35 million a year from commercial sources, needs to prepare young scientists for industry.

Powell says, adding that it's "crazy" for London universities to compete "when the real competition is MIT, Harvard and Stanford"

Tony Jones's job as head of the life sciences part of the business network London First is to stop anxiety about 'who's on top' and focus on collaborations. Academic-industrial partnerships can cut through the competition, he says.

For instance, the London-based global pharmaceutical company GlaxoSmithKline is investing in an imaging centre at Hammersmith Hospital in the city, which is itself affiliated with a number of area universities. North Carolina-based contract research organization Quintiles Transnational is running a clinical trial out of Guy's Hospital that could herald others — to further tap London's large, diverse population of patients for testing drug safety and efficacy.

After years of competing for both prestige and a limited pool of funds, Jones says academic institutions in London seem more willing to get together. For instance, London Higher — the umbrella organization for the city's universities — covers 42 educational institutions, which are pooling their resources to look at issues such as attracting more international talent, especially in terms of students. A bigger population of foreign students could help revitalize the country's knowledge economy, especially in areas like mathematics, physics, and chemistry, where home-grown interest has flagged in recent years. And Genesis, an annual party-turned-meeting of the London Biotechnology Network, is matching academics with companies, and companies with venture capitalists.

Leszek Borysiewicz, deputy rector at Imperial,

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

agrees that there's a problem with recruiting students in core disciplines. "We need people to enter basic science," Borysiewicz says, to take the place of senior investigators approaching retirement. One solution is to tell prospective students that pursuing science and technology degrees will help them land good jobs. Another is to come up with more equitable compensation packages, a route Imperial has taken and Borysiewicz hopes more area institutes will follow.

Improvements in infrastructure can also attract more people to the region, and Imperial has embarked on an ambitious building scheme to do just that. The university is investing US\$179 million this year alone into improving its ageing campus. Imperial is building a new facility for engineering at its South Kensington campus, constructing a developmental biology centre with GlaxoSmithKline on its Hammersmith campus, and putting US\$61 million into St Mary's Hospital in Paddington, primarily to create new facilities for clinical researchers. "That level of expenditure is going to go on for a protracted period," Borysiewicz says.

Borysiewicz adds that although the universities within the 'golden triangle' are essentially competitive, the sky-high cost of the technology needed for the big issues is necessitating strategic collaborations. Imperial and University College London, for instance, have jointly created the London Centre for Nanotechnology. "It would have been very difficult to do it separately," Borysiewicz says.

Collaborations extend beyond infrastructure and joint institutes, to individual research projects. For example, Imperial and St Mary's are piloting a scheme where medical robots would make the rounds to check on patients. And Imperial is exploring similar research

Room to move: with so much space, Cambridge (left, above) is ever ready for an influx of new companies, but the physical limitations of a London site haven't stopped Imperial (right, above) from major expansion.



R. ESTALL/CORBIS/IMPERIAL COLLEGE/I. CHLEBIK

collaborations with long-time rivals Oxford and Cambridge universities.

Robert Ian Lechler, rector of King's College London, is all for cooperation, but doesn't dismiss the power of healthy competition. Using a European football analogy, he sees Imperial — where he was director of immunology — as a team in the premiership, and King's as a championship team aiming to raise itself to that next level.

Lechler is doing some heavy recruiting to improve King's ranking, midway through a plan to bring in 14 research chairs in 12 months, with over US\$3.5 million of funding. Among the newcomers is Graham Lord, a former PhD student of Lechler's whom he has lured away from Harvard. King's is also holding its own in the infrastructure stakes, with a US\$45 million cancer cell-biology building opening at the year's end, and fundraising under way for a clinical neuroscience block.

University College London, too, is making strategic investments, especially in neuroscience. The university's Queen Square campus is becoming a London hub for the discipline, with research hospitals, lab space and the Institute of Cognitive Neuroscience in close proximity. A new building is set to open in 2007.

Although London's status as an academic powerhouse is clearly on the rise, the city still needs to create space for biotech, agree Jones and Powell. Arrow has managed to carve out a space for its 200 or so employees south of the Thames, but most fledgling companies aren't so lucky. That's why London First has helped promote an incubator in Camden.

Cambridge, on the other hand, has an abundance of space — especially after the bursting of the dot-com and telecom bubbles a few years ago. The city was once as close to redundancy-proof as possible; even

casualties of the Glaxo–SmithKline merger soon found work (see *Nature* **414**, 4–5; 2001). But now both landlords and employees are facing a different reality.

A few years ago Boston area biotech company Millennium moved in, only to pull out six weeks later when the company's management retrenched. And a building once occupied by biotech firm Organon is under renovation.

"The demand for occupancy in science and business parks fell off a cliff," says Jeremy Fairbrother, senior bursar at Cambridge and landlord of the university's science park. "But that's life."

That doesn't mean there aren't companies to fill the spaces waiting in the wings. It's just that they're a different breed — smaller, often mixing information technology with hard science and, perhaps more importantly, with market-ready products and services. Fairbrother points to Vectura, which aims to deliver insulin to diabetics through inhalers, or Abcam, a self-described "Amazon of antibodies", providing the proteins over the Internet.

Fairbrother is also seeing people look further afield. They are still finding jobs, but they may have to travel around the edges of the triangle to get to them. One Cambridge telecom worker whom Fairbrother knows keeps his home and family in Cambridge and commutes to London — about two hours of train travel a day. "You have to look a bit wider," he says.

Jeff Solomon, head of Cambridge area biotech network ERBI, says the net number of biotech jobs in the area has remained unchanged over the past few



Eye on recruitment:
Imperial's Leszek
Borysiewicz.

years, but the overall landscape has shifted considerably. "There has not been growth in biotech employment in this region, but neither has there been shrinking." New companies have sprung up where other have withered, and some big players, most notably Celltech — purchased by Brussels-based drug company UCB — have ended up folded into others. Solomon sees that as healthy. "If it's not going on, there's something wrong with your region," he says.

But there are positive signs, too. Although no Cambridge-area companies went public last year, they raised a substantial amount of capital. The region also made strides in establishing alliances among area companies — something that will make the ones that remain stronger in the long run.

One such company, VASTox, is benefiting from its place in the triangle, says company CEO Jon Tinsley. The firm spun out of labs in Oxford, Cambridge and London, raised a significant amount of cash last year and is "trying not to spend it", Tinsley says.

VASTox, which focuses on providing services in biological chemistry, rather than creating its own therapeutics, has been able to recruit scientists from all three cities in the triangle, and beyond "We've actually recruited someone from the United States," Tinsley says.

Tinsley has experience with both academic and industrial jobs within the region. He spent time at Oxford as well as at London's Medical Research Council before joining a company that wasn't as judicious with its cash. Despite the experience, he thinks there are more opportunities in biotech than

A JEWEL IN OXFORD'S CROWN

The Diamond Light Source in Oxfordshire (pictured below) breaks a lot of records. Its US\$668 million price tag is Britain's single biggest investment in research and development infrastructure. The synchrotron's beam lines, designed to probe the molecular nature of both proteins and non-biological materials, will be among the world's most powerful when it comes online in 2007. And the 300 highly skilled employees needed to run it represent one of the country's most ambitious scientific recruitment exercises.

But the location of the facility, in expensive Oxfordshire rather than near the older synchrotron in Daresbury, Cheshire, is also a source of serious and lingering controversy (see *Nature* **402**, 451; 1999).

Perhaps as a result of the furore over location, only a minority of employees from the Cheshire facility have chosen to make the move south. Louise Johnson (right), Diamond's life-science director in Daresbury, says that so far, only two of the ten scientists in her group have committed to moving with her. She says that figure is fairly representative among other Diamond group leaders.

Johnson is, however, working hard to make the Oxfordshire location an attractive one for



researchers. She's able to offer tenure-track, long-term contracts. It's a substantial facility, with 2,000 square metres of research space available in the doughnut-shaped building. And Britain's Medical Research Council is funding a research complex nearby that will help technicians do their own research, rather than solely assisting others. "People are excited about being involved in a large project," says Johnson. **P.S.**



academia now, and sees the emergence of companies like London-based Arrow and Antisoma, both inching towards getting products to market, as signs that the triangle can be golden for biotech employees.

Lin Bateson, operations manager of the Oxfordshire Bioscience Network, says VASTox is typical of a handful of companies thriving at each point of the triangle. Last year, a few in each area attracted a disproportionate amount of venture capital: five companies took 68% of the US\$250 million that flowed into Oxfordshire, Bateson says.

Although a few strong contenders are emerging, Bateson worries about younger companies. The government is lending a hand to tech transfer, but providing little support to spin-outs.

Meanwhile, just as universities are working to prepare students and postdocs for the time when the baby boomers retire from their academic posts, they must also set up young scientists for industrial jobs, says Nigel Thrift, now head of Oxford's division of life and environmental sciences and soon to be provost chancellor of research. The university is more reliant on industry money now, getting US\$36 million a year from companies, compared with US\$28 million from government and charities.

Some interactions are clearly beneficial. For example, Google is involved in Oxford's main library. And the university's chemistry department has fed many scientists, particularly medicinal chemists, into the pharmaceutical industry. But the university is upping its ante in the life sciences by

LEADING INDICATORS: SOUTHEAST ENGLAND

Funding

Domestic research and development funding: 1.86% of GDP.

The UK Medical Research Council alone is awarding more than US\$300 million in grants for 2004–5 — a disproportionate amount of which stays in southeast England. London hosts more than a third of publicly funded research in Britain. The Wellcome Trust is the world's largest biomedical research foundation, giving out more than US\$700 million a year, much of which goes to researchers in London, Cambridge and Oxford.

Biotech companies

Britain ranks fourth globally and second in Europe in number of companies.

Biotech clusters

London has more than 100 biotech companies; Cambridge has about 200; Oxford, 67.

Park and go:
Cambridge
Science Park
stalwarts such
as Napp
Pharmaceuticals
have survived
huge changes.

building the Oxford Centre for Cellular and Molecular Biology. The building will hold 550 researchers and should include many new hires, not just a reshuffling of existing faculty, Thrift says. The same goes for the Diamond Light Source, a synchrotron which uses high-energy beams of light to probe the molecular structure of many proteins and nonbiological materials. The project is hiring many new people, as staff from the older synchrotron at Daresbury, Cheshire, have decided not to make the move south (see 'A jewel in Oxford's crown'). The university also has a programme that trains physical scientists for life-science work.

Raymond Dwek, head of Oxford's biochemistry department, says the university has one of the largest collection of academic biochemists in the world — 800 in total, including about 100 professors. Dwek, who has spent more than 40 years at the university, wants his legacy to be the addition of 20 professors next year and the building of a US\$154 million facility. "At a time when departments are contracting, we are expanding," Dwek says.

His successor, Kim Nasmyth, former head of the IMP in Vienna, will join and add some molecular biology strength to the chemical skills the department possesses. Nasmyth will build an institute of chromosome biology, while Dwek, a founder of Oxford Glycosciences, will turn his strengths in sugar molecules to nurturing Oxford's institute of glycobiology.

Dwek, like Borysiewicz at Imperial, Lechler at King's, and Fairbrother at Cambridge, is almost gleeful when discussing his plans to outdo his peers around the triangle. But he has a positive perspective on the benefits of such competition. "As long as the rivalry helps the science, helps the public, helps accountability, that's OK," he says. ■

Paul Smaglik is *Naturejobs* editor

WEB LINKS

University of Oxford
 ▶ www.ox.ac.uk
 University of Cambridge
 ▶ www.cam.ac.uk
 Imperial College London
 ▶ www.imperial.ac.uk
 University College London
 ▶ www.ucl.ac.uk
 King's College London
 ▶ www.kcl.ac.uk
 Wellcome Trust
 ▶ www.wellcome.ac.uk
 Sanger Centre
 ▶ www.sanger.ac.uk

BioLondon
 ▶ www.biolondon.org.uk
 Oxfordshire BioScience
 Network
 ▶ www.oxfordshirebioscience.com
 ERBI
 ▶ www.erbi.co.uk
 South East England
 Development
 Agency
 ▶ www.seeda.co.uk
 Diamond Light Source
 ▶ www.diamond.ac.uk

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

P. WARD/CORBIS

What's expected of us

It's a tough choice...

Ted Chiang

This is a warning. Please read carefully.

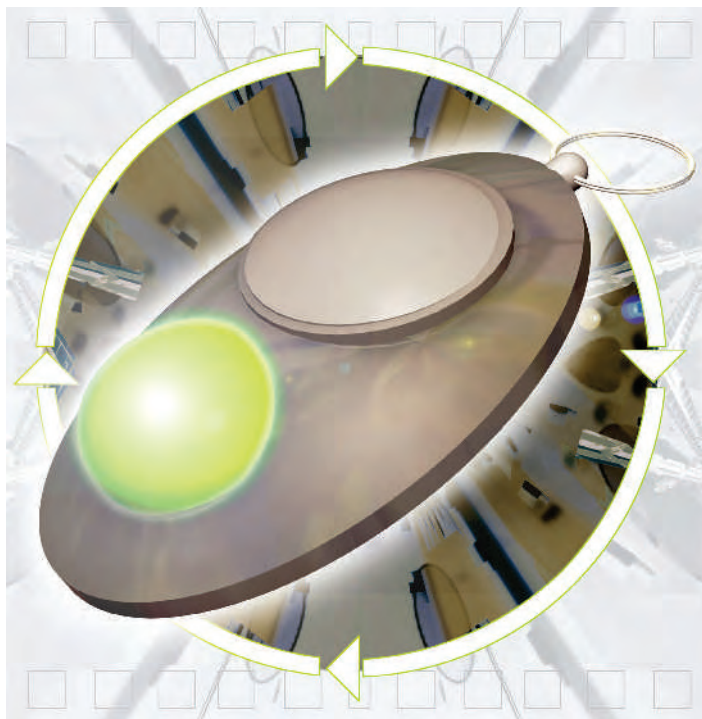
By now you've probably seen a Predictor; millions of them have been sold by the time you're reading this. For those who haven't seen one, it's a small device, like a remote for opening your car door. Its only features are a button and a big green LED. The light flashes if you press the button. Specifically, the light flashes one second *before* you press the button.

Most people say that when they first try it, it feels like they're playing a strange game, one where the goal is to press the button after seeing the flash, and it's easy to play. But when you try to break the rules, you find that you can't. If you try to press the button without having seen a flash, the flash immediately appears, and no matter how fast you move, you never push the button until a second has elapsed. If you wait for the flash, intending to keep from pressing the button afterwards, the flash never appears. No matter what you do, the light always precedes the button press. There's no way to fool a Predictor.

The heart of each Predictor is a circuit with a negative time delay — it sends a signal back in time. The full implications of the technology will become apparent later, when negative delays of greater than a second are achieved, but that's not what this warning is about. The immediate problem is that Predictors demonstrate that there's no such thing as free will.

There have always been arguments showing that free will is an illusion, some based on hard physics, others based on pure logic. Most people agree these arguments are irrefutable, but no one ever really accepts the conclusion. The experience of having free will is too powerful for an argument to overrule. What it takes is a demonstration, and that's what a Predictor provides.

Typically, a person plays with a Predictor compulsively for several days, showing it to friends, trying various schemes to outwit the device. The person may appear to lose interest in it, but no one can forget what it means — over the following weeks, the



implications of an immutable future sink in. Some people, realizing that their choices don't matter, refuse to make any choices at all. Like a legion of Bartleby the Scriveners, they no longer engage in spontaneous action. Eventually, a third of those who play with a Predictor must be hospitalized because they won't feed themselves. The end state is akinetic mutism, a kind of waking coma. They'll track motion with their eyes, and change position occasionally, but nothing more. The ability to move remains, but the motivation is gone.

Before people started playing with Predictors, akinetic mutism was very rare, a result of damage to the anterior cingulate region of the brain. Now it spreads like a cognitive plague. People used to speculate about a thought that destroys the thinker, some unspeakable lovecraftian horror, or a Gödel sentence that crashes the human logical system. It turns out that the disabling thought is one that we've all encountered: the idea that free will doesn't exist. It just wasn't harmful until you believed it.

Doctors try arguing with the patients while they still respond to conversation. We had all been living happy, active lives before, they reason, and we hadn't had free will then either. Why should anything change? "No action you took last month was any more freely chosen than one you take today," a doctor might say. "You can still behave that way now." The patients invari-

ably respond, "But now I know." And some of them never say anything again.

Some will argue that the fact the Predictor causes this change in behaviour means that we *do* have free will. An automaton cannot become discouraged, only a free-thinking entity can. The fact that some individuals descend into akinetic mutism whereas others do not just highlights the importance of making a choice.

Unfortunately, such reasoning is faulty: every form of behaviour is compatible with determinism. One dynamic system

might fall into a basin of attraction and wind up at a fixed point, whereas another exhibits chaotic behaviour indefinitely, but both are completely deterministic.

I'm transmitting this warning to you from just over a year in your future: it's the first lengthy message received when circuits with negative delays in the megasecond range are used to build communication devices. Other messages will follow, addressing other issues. My message to you is this: pretend that you have free will. It's essential that you behave as if your decisions matter, even though you know that they don't. The reality isn't important: what's important is your belief, and believing the lie is the only way to avoid a waking coma. Civilization now depends on self-deception. Perhaps it always has.

And yet I know that, because free will is an illusion, it's all predetermined who will descend into akinetic mutism and who won't. There's nothing anyone can do about it — you can't choose the effect the Predictor has on you. Some of you will succumb and some of you won't, and my sending this warning won't alter those proportions. So why did I do it?

Because I had no choice. ■
Ted Chiang is an occasional writer of science fiction. His work can be found in his collection *Stories of Your Life and Others*, published by Pan Macmillan.

JACEY